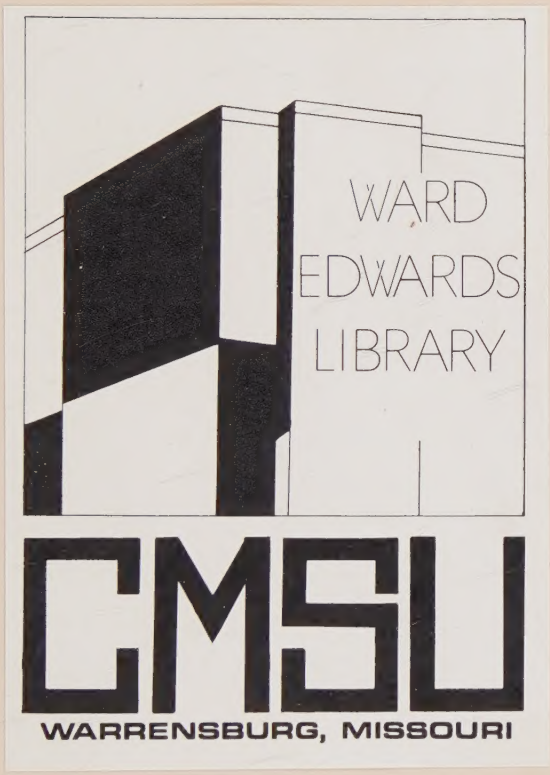
LB
1131
K45

Columbia University

Contributions to Education

Teachers College Series

State Teachers College
Warrensburg, Mo.
Class 371.26 Book K523
Accession 77294


WARD
EDWARDS
LIBRARY
CMSU
WARRENSBURG, MISSOURI

THE IMPROVEMENT OF MEASUREMENT THROUGH CUMULATIVE TESTING

AN EMPIRICAL STUDY OF TWO HUNDRED ELEMENTARY SCHOOL CHILDREN OVER A PERIOD OF FOUR YEARS

BY

NOEL KEYS, PH. D.

ASSOCIATE PROFESSOR OF EDUCATION, UNIVERSITY OF CALIFORNIA

TEACHERS COLLEGE, COLUMBIA UNIVERSITY
CONTRIBUTIONS TO EDUCATION, NO. 321

PROPERTY OF
TEACHERS COLLEGE
WARRENSBURG
CENTRAL MISSOURI
STATE UNIVERSITY
Warrensburg,
Missouri

BUREAU OF PUBLICATIONS
Teachers College, Columbia University
NEW YORK CITY
1928

Copyright, 1928, by Teachers College, Columbia University

PROPERTY OF
TEACHERS COLLEGE
WARRENSBURG MO.

LB
1131
K45

ACKNOWLEDGMENTS

For the means of effecting the present study, the writer's thanks are due primarily to Director Eugene A. Nifenecker and Assistant Director William A. McGrath, of the Bureau of Reference, Research and Statistics of the New York City Board of Education. Except for the records of the three-year testing program which these gentlemen so generously placed at his disposal, the task would have been much too great for his resources. Special acknowledgment is also due to Principals Mary C. Bergen of Public School 59, Manhattan, Jacob Theobald of Public School 165, Manhattan, and Charles B. Jameson of Public School 58, Queens, Greater New York. Only the whole-hearted coöperation of these principals and their able assistants made possible the final steps in the investigation.

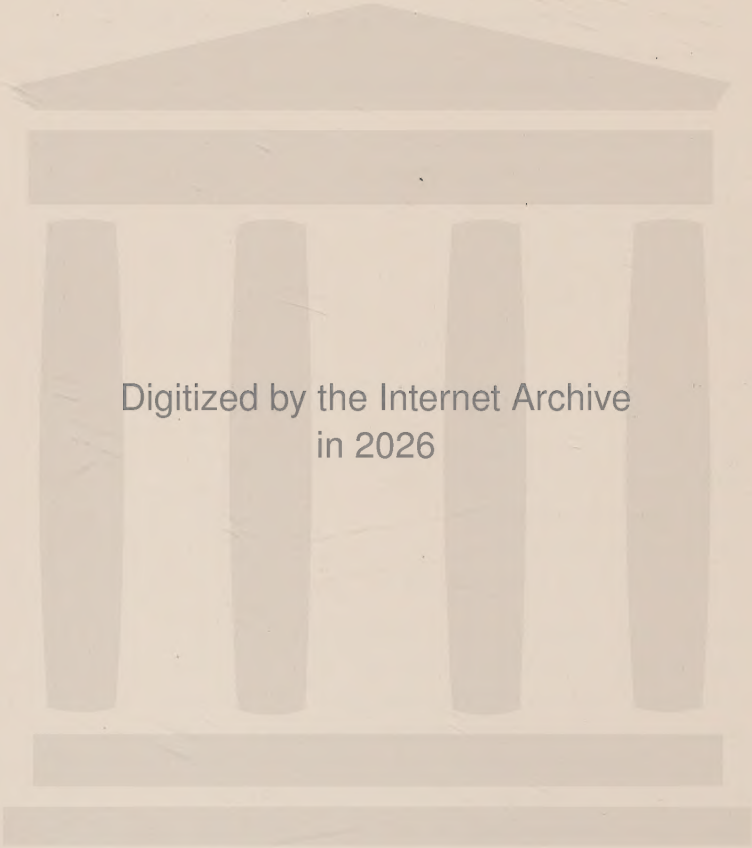
The writer wishes to express his grateful appreciation of the counsel and encouragement received from Professors A. I. Gates, W. A. McCall, R. Pintner, P. M. Symonds, and B. D. Wood, of Teachers College, Columbia University, under whose helpful direction the study was made, and his deep indebtedness to his mother for prolonged assistance in the compilation of the data and the ensuing computations.

N. K.

77294

CONTENTS

CHAPTER		PAGE
I.	The Problem Defined	1
	Previous Studies Throwing Light Upon the Effect of Time on Size of Correlations	3
	Summary and Statement of Present Problem	7
II.	Nature and Sources of Data	10
	Character of the Constant Group	11
	Description of Tests Used	17
	Method of Scaling Tests	18
	Method of Equating Forms	22
	Validity of the Scaling	25
	The Criterion of Measurement	28
III.	Results: The Influence of Time Upon Scores	30
	Intercorrelations Obtained	32
	Reliability of the Coefficients	35
	Influence of Time on Correlations	36
	Coefficients of Reliability	42
	Curves of Improvement	45
	Influence of Time Upon Variability	47
IV.	Results: The Advantages of Cumulation	51
	The Multiple Correlations Obtained	53
	Correlations with the Unweighted Mean	55
	Limits to the Advantages of Simple Averaging	56
	Average Errors of Estimate	60
	Comparative Advantages of Different Types of Score	64
	The Improvement Evaluated	69
V.	Summary and Conclusions	74
	Bibliography	80


Digitized by the Internet Archive
in 2026

https://archive.org/details/improvementofmea0000noel

TABLES

TABLE		PAGE
A.	Coefficients of Correlation for Stanford-Binet IQ's	4
B.	Mean Intercorrelations for Raw Scores on Various Tests over the Intervals Indicated (After Gates and LaSalle)	7
I.	Sex and School Derivation of the Constant Group	12
II.	Grade Location of the Constant Group at Times of Testing	12
III.	Quotients of the Two Hundred Pupils, Indicating Distribution of Ability in the Constant Group	14
IV.	Brightness Scores of the Two Hundreds Pupils as in Table III	15
V.	Standard Deviation of Ages and Test Scores of the Constant Group in May, 1924, as Contrasted with Those for Pupils of the Nearest Corresponding Year-Grade in Twelve New York City Schools Combined	16
VI.	Extract from Tables Employed in Determining T Scores for Thorndike-McCall Reading Form 1 by Percentile Equation with Form 2, Already T Scaled	24
VII.	Constancy of Intelligence Quotient in Constant Group when Based on Present Scaling and Local Norms as Contrasted with Those of the National Intelligence Test Manual	28
VIII.	Coefficients of Correlation for Raw Scores on Tests	32
IX.	Coefficients of Correlation for T Scores	33
X.	Coefficients of Correlation for Quotients	33
XI.	Coefficients of Correlation for Brightness Scores	34
XII.	Probable Errors for Coefficients of Correlation Based on a Group of Two Hundred	35
XIII.	Mean Coefficients of Correlation Over Varying Time Intervals	37
XIV.	Corresponding Coefficients of Correlation Between Tests of Equal Reliability Given to Like Groups After Different Intervals of Time	40
XV.	Coefficients of Reliability for Tests of the Present Study Expressed as for Groups of the Variability of Twelve-Year-Old Children of the Schools Concerned, and Corrected for the Effects of Time Elapsed Between Tests	43
XVI.	Curves of Improvement of the Six Tests and Educational Batter Score, as Measured by Mean T Scores of Successive Testings	46
XVII.	Standard Deviations of Scores of the Constant Group, Showing Tendency to Increase with Successive Testings	49

TABLE PAGE

XVIII. Relative Variability in Terms of the Pearson Coefficient ($V = 100\sigma/M$) as Based on T Scores of the Constant Group on Successive Testings 50

XIX. Multiple Correlations Between Actual 1926 Performance and Predictions Based Upon Two or More Previous Testings Using Raw Scores in the Original Test Units 52

XX. Multiple Correlations Between Actual 1926 Performance and Predictions Based Upon Two or More Previous Testings Using Scores in Form of Quotients 53

XXI. Multiple Correlations Between Actual 1926 Performance and Predictions Based Upon Two or More Previous Testings Using Brightness Scores 54

XXII. Mean Correlations with 1926 57

XXIII. Multiple Correlations for a Hypothetical "Type Test" Given at One-Year Intervals Over a Period of Five Years..... 59

XXIV. Mean Error of Estimate, or Average Difference Between the Standing of Individual Pupils in 1926 as Measured and as Predicted by Simple Averaging of Their Standing in Previous Years 62

XXV. Correlations Between Final Scores and Those Predicted with and without the Use of Initial Measures of Brightness (Quotients or B Scores) 67

XXVI. Summary Table Indicating Degree of Improvement Attained Through Cumulation of Test Results for Various Tests and Types of Score with and without the Use of Weighting 70

THE IMPROVEMENT OF MEASUREMENT THROUGH CUMULATIVE TESTING

AN EMPIRICAL STUDY OF TWO HUNDRED ELEMENTARY SCHOOL CHILDREN OVER A PERIOD OF FOUR YEARS

CHAPTER I

THE PROBLEM DEFINED

The pressing need for an increased reliability in the majority of our educational measurements is quite generally recognized. Huffaker [1] and Herring,[2] taking the standard deviation of unselected twelve-year-olds on the Stanford Revision of the Binet Scale as 18 months of mental age, have used the Standard Error of Measurement ($\sigma\sqrt{1 - r_{12}}$) to compute the percentage of pupils who will be mismeasured by stated amounts, given tests of varying degrees of reliability. Thus, if we adopt as a minimal standard the requirement that not over 12 per cent of twelve-year-olds be placed more than twelve months of mental age from their true status, Huffaker has shown that our test must show a self-correlation of .80; if we demand that only 7 per cent be so mismeasured, the correlation must be .85; while to reduce the figure to 3 per cent, the test must yield a correlation of .90 or better when applied to unselected twelve-year-olds. Although the best existing measures of intelligence and achievement, both individual and group tests, can satisfy the last requirement, there is a greater number which cannot. Yet many of the latter tests possess qualities of convenience, diagnostic discrimination, and the like, which render their retention highly desirable, provided some means can be found of increasing their reliability.

The statistics of measurement suggest two ways in which this might be done.

[1] Huffaker, C.L. "Reliability of Measurement by Group Tests of Mental Ability." *Journal of Educational Psychology*, XVI: 493-97, 1925.

[2] Herring, J.P. "Reply to Huffaker's Criticism." *Journal of Educational Psychology*, XVI: 498-99, 1925.

1. By increasing the length of the tests until the requisite reliability is secured. While doubtless the most thoroughgoing and satisfactory solution, this would entail an actual reconstruction and restandardization of the tests themselves, and so is of little avail to the practicing schoolman or teacher seeking to utilize existing instruments.

2. By giving the same or parallel forms of a test two or more times and averaging the scores. The result of this, so far as statistical factors are concerned,[3] would be the same as from the use of a longer test and could be accomplished with the tests as they stand. On the other hand, the time and expense required would be as great as or greater than that for the single, longer test.

It frequently happens, however, that reasons of administration, supervision of instruction, group experimentation, and the like make it desirable to retest at intervals of months or years, large groups of school children in respect to certain abilities. Recurrent surveys of this type are an increasingly common feature in modern administrative practice. For such purposes group tests of only mediocre reliability are often used, and properly so. The question arises whether it may not be possible to utilize the results of such repeated measurements with ordinary group tests to increase the reliability of the latest measurement to the point where it may be used for evaluating the work of the individual pupil.[4]

To this problem the present study is addressed: How and to what extent can the cumulative results of repeated measurements be used to increase the reliability of individual scores, while reducing the necessity for costly and time-consuming testings on any one occasion? The answer will, of course, turn largely upon the extent to which lapse of time serves to invalidate the rating of an individual in different abilities and functions. Various investigations have served to shed some light upon this issue.

[3] Garrett, H.E. *Statistics in Psychology and Education,* pp. 269-70. Longmans, Green, 1926.

[4] This would, of course, imply a system of cumulative individual records whereby the teacher or supervisor would have available for each child the record of his previous performances on a given test, to be averaged with the results of the latest application. But such records are amply justified on other grounds. Properly devised, they should be of the greatest value, both as an aid to individual diagnosis and as the source of data for a wide variety of personnel researches.

PREVIOUS STUDIES THROWING LIGHT UPON THE EFFECT OF TIME ON SIZE OF CORRELATIONS

Studies of retests of intelligence using individual tests, and particularly the Binet-Simon, are both too numerous and too well known to justify extensive review in this place. Apart from the monumental investigations of Kuhlmann [5] with his 1912 and 1917 revisions and Terman [6] with the Stanford Revision, Baldwin,[7] Dickson,[8] Rugg,[9] Poull,[10] Madsen,[11] Gray and Marsden,[12] and Garrison and Robinson [13] are but a few of the many who have reported significant results with retests and the constancy of the IQ. While numerical findings differ widely according to the variability of the groups tested and the length of the interval between tests, there appears a fair consensus that, as regards the Stanford Binet IQ, correlations remain relatively high over a period of years, though declining steadily with lapse of time, and that the mean arithmetic difference on two successive testings is likely to be between 4 and 6 IQ points. Representative of the less satisfactory results are Berry's [14] findings for 390 pupils in Detroit "Special Classes." Over an interval of eleven months the correlation obtained was .74, over twenty-three months .67, and over thirty-five months, but .56.

Baldwin and Stecher have furnished one of the most prolonged surveys, showing intercorrelations for six annual examinations of 143 normal and superior children, 32 of whom were tested all six times. The results reported by them are shown in Table A.

[5] Kuhlmann, F. "The Results of Repeated Examinations of 639 Feeble-Minded over a Period of Ten Years." *Journal of Applied Psychology*, V: 195-224, 1921.

[6] Terman, L.M. *The Intelligence of School Children.* Houghton Mifflin, 1919.

[7] Baldwin, B.T. and Stecher, L.I. *Mental Growth Curve of Normal and Superior Children Studied by Means of Consecutive Intelligence Examinations.* University of Iowa Studies in Child Welfare, Vol. II, No. 1, 1922. Contains good bibliography.

[8] Dickson, V.E. *Mental Tests and the Classroom Teacher*, pp. 67-68. World Book Co., 1923.

[9] Rugg, L.S. "Retests and the Constancy of the IQ." *Journal of Educational Psychology*, XVI: 341-43, 1925.

[10] Poull, L.E. "Constancy of the IQ in Mental Defectives." *Journal of Educational Psychology*, XII: 366-72, 1921.

[11] Madsen, I.N. "Some Results with the Stanford Revision." *School and Society*, 19:559-62, 1924.

[12] Gray, P.L. and Marsden, R.E. "The Constancy of the IQ—Further Results." *British Journal of Psychology*, XV: 169-73, 1924-25.

[13] Garrison, S.C. and Robinson, M.S. "A Study of Retests." *Journal of Educational Research*, XI: 190-96, 1925.

[14] Berry, C.S. "The Intelligence Quotients of Mentally Retarded Children." *School and Society*, 17: 723-29, 1923.

TABLE A

COEFFICIENTS OF CORRELATION FOR STANFORD-BINET IQ'S [15]

Examination Number	1	2	3	4	5
2	.850 ± .031				
3	.738 ± .051	.846 ± .031			
4	.779 ± .044	.802 ± .040	.910 ± .019		
5	.817 ± .037	.815 ± .037	.839 ± .033	.918 ± .017	
6	.812 ± .038	.751 ± .049	.796 ± .041	.866 ± .028	.944 ± .012

On combining and averaging certain of the above correlations, two facts become apparent:

1. Coefficients decrease with increasing time-interval, though the observed relation is somewhat irregular.

Length of Interval	Number of Correlations Averaged	Mean *r*
One year	5	.894
Two years	4	.811
Three years	3	.797
Four and five years	3	.793

2. Later correlations exceed the earlier. The coefficients for one-year intervals, in order from first to last, read: .850, .846, .910, .918, .944. The progression here is more clearly defined than in the case of differing time-intervals, and strongly suggests an increasing reliability with increasing familiarity with the test (see Chapter III).

Turning to studies not confined to Binet test results, we find the most comprehensive treatment of the effect of time upon correlations to be that of Woodyard,[16] who has assembled correlations in a wide variety of functions, both capacity and achievement, for groups ranging from fourth grade classes to middle-aged adults. The conclusion reached was that

> This factor [lapse of time] is of small importance in estimating the total variability of an individual. The progressive variability with increasing lapse of time is very slight if it truly exists at all. Very little more difference can be found after a year than after a week, and little more after some days than after some hours. . . . In determining the reliability of a testing instrument . . . the repeat test may be given . . . a year later with but small difference in the results.

[15] From Baldwin, B.T. and Stecher, L.I. "Additional Data from Consecutive Stanford-Binet Tests." *Journal of Educational Psychology*, XIII: 556, 1922.

[16] Woodyard, Ella. *The Effect of Time Upon Variability*. Teachers College, Columbia University, Contributions to Education, No. 216, 1926.

This conclusion is immediately followed by a qualifying statement, as quoted below:

> The effect of a time interval is somewhat more uncertain upon children than upon adults, the results for the children in Grades 4 to 8 showing lessening variability among the older children, while in the numerous tests upon seventh grade pupils, of the 15 sets of data shown, 9 show results favoring decreased coefficients of correlation after a time interval, 4 show increased coefficients, and 2 show no change.

It should be noted, however, that the intervals studied by Woodyard did not exceed one year, and the conclusions reached would scarcely be intended to apply to periods longer than that.

In 1917, Jones [17] reported results of retests on 203 Cincinnati children at the ages of 14, 15, 16, and 17, using cancellation, digit-symbol substitution, immediate memory, sentence completion, and opposites. Individuals were scored merely according to the decile in which they fell, and these scores were then averaged. His correlations by this method of decile status in age groups ranged from .69 to .76, but showed no tendency whatever to decrease with lapse of time. He did, however, remark (p. 84) an increase in standard deviations with age, which he attributed to "the varying conditions of work in Cincinnati after the age of sixteen and perhaps other factors" operating "to make the good slightly better and the poor relatively poorer than they were at the age of fourteen."

Garrison and Robinson [18] retested 131 pupils in Grades 3 to 8 of the Peabody Demonstration School with the National Intelligence Test A and the Stanford Binet, each given at ten months and at twenty months after the first testing. The correlations between test scores for this widely diversified and "test-wise" group were as follows:

	N. I. T.–A		Stanford Binet	
Test number	1	2	1	2
2	.90		.88	
3	.91	.93	.91	.92

It will be noted:

1. That, as with Jones and Woodyard, the foregoing correlations show no tendency to decrease with time.

[17] Jones, E.S. *Influence of Age and Experience on Correlations Concerned with Mental Tests.* University of Chicago Press, 1917.

[18] Garrison, S.C. and Robinson, M.S. "A Study of Retests." *Journal of Educational Research,* XI: 190-96, 1925.

2. That those for the National Intelligence Test are fully as high as those for the Stanford Binet.

Of interest also are the findings that the intelligence rating obtained by averaging results of the three testings yielded correlations with educational tests and with teachers' judgments of .79 and .76 respectively for the National, as compared with only .58 and .72 for the Binet.

Pintner has at different times [19] reported correlations obtained with his Non-Language Intelligence and Educational Survey Tests over periods of a few days, three years, and four years, respectively. The smallness of the numbers and the fact that entirely different groups of children were involved in the case of the different intervals make it impossible to judge the precise effect of time upon the agreement. The one conclusion which may safely be drawn from the data is that, as between these two tests and as applied to deaf children at least, educational indices are fully as constant and reliable over considerable periods as are mental.

Brooks [20] retested 67 children in twenty different functions after two intervals of one year each. He concluded that the relative status of the individual in the more complex and "higher" functions, such as reading, arithmetic, spelling, sentence-completion, composition, and the like, is much more constant than in such simpler abilities as number checking, handwriting, auditory memory-span, and the like.

The most pertinent and thoroughgoing study reported with achievement tests is probably that of A. I. Gates.[21] Together with Miss LaSalle, Gates tested the same 78 third to sixth grade pupils of Scarborough School with a battery of educational tests, given at four-month intervals over two school years. Correlations were obtained for five subjects over intervals ranging

[19] Pintner, R. "Group Tests after Several Years." *Journal of Educational Psychology*, XVI: 391-95, 1925.

Pintner, R. "Results Obtained with the Non-Language Group Test." *Journal of Educational Psychology*, XV: 42-48, 1924.

[20] Brooks, F.D. *Changes in Mental Traits with Age: Determined by Individual Retests.* Teachers College, Columbia University, Contributions to Education, No. 116, 1921.

[21] Gates, A.I. and LaSalle, Jessie. "The Relative Predictive Value of Certain Intelligence and Educational Tests, Together with a Study of the Effect of Educational Achievement upon Intelligence Test Scores." *Journal of Educational Psychology*, XIV: 517-39, 1923.

Also Gates and LaSalle. "A Study of Writing Ability." *Journal of Educational Psychology*, XV: 205-16, 1924.

TABLE B

MEAN INTERCORRELATIONS FOR RAW SCORES ON VARIOUS TESTS OVER THE INTERVALS INDICATED (After Gates and LaSalle)

Interval in Months	Reading Comprehension (Thorndike-McCall and Burgess)	Reading Rate (Courtis)	Arithmetic (Woody, four forms)	Spelling (50 words from Ayres list)	Handwriting Speed and Quality (Thorndike)
4	.79 ± .01	.80 ± .01	.91 ± .01	.90 ± .00	.90 ± .01
8	.77 ± .01	.77 ± .01	.88 ± .01	.90 ± .01	.88 ± .01
12	.75 ± .01	.73 ± .02	.89 ± .01	.87 ± .01	.88 ± .01
16	.76 ± .00	.73 ± .03	.88 ± .01	.88 ± .01	.86 ± .01
20	.75 ± .00	.74 ± .00	.84 ± .00	.88 ± .00	.85 ± .02

from four to twenty months. As assembled by the writer in Table B, these correlations reveal a small but fairly consistent decline as the intervals increase. Similar results were obtained by G. S. Gates,[22] even with simple functions, such as color-naming and tapping-rate. It should, however, be noted that the Scarborough correlations were based entirely upon raw scores, and were secured from a composite group of children in grades ranging from third to sixth. It is only reasonable to suppose that children at such widely different stages of learning would be advancing at very different rates, and that this in itself would be reflected in decreasing correlations with passage of time.

SUMMARY AND STATEMENT OF PRESENT PROBLEM

I. Our survey of previous literature has shown that:

1. Studies, except with Binet tests, have been few and heterogeneous, with results accordingly conflicting. They do not afford the data for an answer to our problem, and even the following conclusions are but tentative.

2. The preponderance of evidence appears to point a tendency for correlations to decrease with time, even when variations in age or level are allowed for by such expedients as the IQ (Baldwin), but the presence of such decline is doubtful in the case of persons above the age of fourteen (Jones). With groups of the homogeneity of a typical school grade, the influence of time, even upon correlations for raw scores, is almost negligible up to intervals of one year (Woodyard).

[22] Gates G.S. *Individual Differences as Affected by Practice.* Columbia University, Archives of Psychology, Vol. 58, 1922.

3. Nothing in these results would indicate that objectively measured performance in elementary school subjects, such as reading, spelling, and arithmetic, is any less constant than scores on tests of general intelligence, so-called (Pintner, Brooks, Gates).

4. For pupils in a given school, scores and IQ's on the better group tests of intelligence, such as the National Intelligence Test, yield correlations which are, up to two years at least, as high as or higher than those obtained for our best individual test, the Stanford Binet.

5. The two studies involving consecutive retests of educational achievement are confined in each case to the pupils of a single private school, where homogeneity of conditions might make educational achievement almost a direct measure of intelligence. In none of the cases noted is the performance of the same pupil on periodical tests of educational achievement reported over a period greater than two years.

II. The present study contrasts with the foregoing in that:

1. It is addressed to the problem of determining the value of cumulative measurements with standard tests over considerable intervals as a means of increasing the reliability of measurement and prediction for the individual.

2. It deals, not with individual tests of high reliability such as the Stanford Binet, but with group tests of intelligence and school achievement, ranging in reliability from the most satisfactory to others of minimal value for individual prediction.

3. The periodical testings are extended over an interval of four calendar years, a total class range of nine half-grades.

4. The scores traced are those of two hundred children in three different public schools, all of whom have taken every test in the series. The constancy and size of this group give to successive means and standard deviations a significance otherwise unobtainable.

5. The testing, except for the final or criterion measurement, is done, not by expert psychologists or measurement specialists, but by regular teachers from the schools concerned, under conditions reproducible in any good, modern public school.

6. The validity of the norms for the several types of test and scale unit under consideration has been rendered as comparable as possible by a restandardization and scaling of all tests accord-

ing to a uniform procedure. The method employed is one which could readily be used in any large school system for the development of additional tests and parallel forms as needed, year after year.

III. In considering the problem of the utility of cumulating records of group test results, this study will seek to throw light upon the following subordinate issues:

1. How and to what extent does lapse of time affect the rating obtained upon an individual over periods of more than one year?
2. Is there a notable difference in this respect between performance on tests of school achievement and on group tests of intelligence, so called?
3. To what extent will averaging with test records secured long previously improve upon the prediction derived from the most recent testing?
4. How far into the past may results be drawn upon to advantage for such averaging?
5. Must the successive measurements be weighted in averaging, as implied in ordinary multiple correlation, or will a simple, unweighted average yield substantially as good results?
6. As between the more commonly employed types of relative score, does one lend itself to notably better advantage for purposes of cumulation and prediction than the other?

CHAPTER II

NATURE AND SOURCES OF DATA

For an investigation of the problems here involved Grades 3 to 8 inclusive offer certain peculiar advantages. It is at this level that we have the most numerous tests available for subjects of study pursued throughout the period and applicable at all stages of advancement. Moreover, this is a period of rapid intellectual growth, which might be expected, in view of individual differences, to give rise to marked shifts in relative position within comparatively short spaces of time. Conclusions as to constancy of status and achievement based on such a group would promise to hold *a fortiori* for stages above elementary school.

The present attack was made possible through the courtesy of Directors Eugene A. Nifenecker and William A. McGrath, of the Bureau of Reference, Research and Statistics, New York City Schools. These men placed at the disposal of the writer the records from a three-year survey made under their direction with a battery of intelligence and achievement tests given to more than ten thousand pupils in Grades 4 to 8 of twelve large public schools of Greater New York. These tests had been given three times at intervals of approximately one and one-half years, in the weeks of December 4, 1922, May 26, 1924, and December 14, 1925.

A tracing of individual records revealed that in the three schools which continued to offer a full eight-grade course there remained in 1926 more than two hundred pupils who had taken all of the tests in the three surveys. Through the generous coöperation of the principals and staffs of these schools, the writer was enabled to assemble these pupils and administer a fourth and final test with the same battery. This was given in the week of December 13, 1926, one year after the latest previous testing and four years after the first. In order to complete the data needed for a restandardization of the tests, as

explained later, the test forms previously used on Grades 4 to 8 were also given to all third grade classes, more than six hundred pupils, in the three schools involved.

Both the giving and the scoring of tests in the previous surveys had been done by regular teachers of the schools concerned. These teachers were trained and supervised by the Bureau of Reference, Research and Statistics. In order that the final, or criterion, testing might be of the utmost possible uniformity for all groups, the 1926 tests were given under the personal supervision of the writer by twelve picked examiners, specializing in measurements at Teachers College, Columbia University.

Testing conditions were favorable, and the enthusiastic coöperation of the pupils was not to be questioned. Indeed, they seemed almost unduly motivated by the consciousness that they had been singled out in some way as specially representing their several schools.

CHARACTER OF THE CONSTANT GROUP

Completion of the final testing showed just 201 individuals with unbroken records for all tests in the four surveys. For statistical convenience, one of these was discarded, leaving the two hundred pupils to be designated hereafter as the "constant group." Since all correlations are based upon results for this group as a whole, without attempt to treat separately different sexes, grades, and so on, a careful consideration of its make-up is in order.

The three schools from which these pupils were drawn differ considerably in several respects:

P. S. 58, Queens, is a coeducational school in a suburban, industrial section of Long Island. Nationalities are mixed, with American and Italian predominating. The children are of about average ability as compared with New York City in general, and the population is somewhat less migratory than most. This accounts in part for the larger number of pupils from this school to be included in the constant group.

P. S. 59, Manhattan, is, in the upper grades, a school for girls only. Located on the central East Side, its pupils are of Polish, Italian, Irish, German, and other descent. The average intelligence is probably below that of the city at large.

P. S. 165, Manhattan, is, in its upper grades, a school for boys only. It is located in a well-to-do residential section, not far from Columbia University. The pupils are largely Jewish, and of better than average intelligence.

Among these schools the two hundred were distributed as follows:

TABLE I

SEX AND SCHOOL DERIVATION OF THE CONSTANT GROUP

School		Boys	Girls	Total
P. S. 58	Queens	66	39	105
P. S. 59	Manhattan	0	70	70
P. S. 165	Manhattan	25	0	25
Total		91	109	200

Distribution of the group among the grades may be seen from Table II. It will be noted that on each of the last three test-

TABLE II

GRADE LOCATION OF THE CONSTANT GROUP AT TIMES OF TESTING

Grade	Number of Pupils			
	Dec., 1922	May, 1924	Dec., 1925	Dec., 1926
Low 4	90			
High 4	59	3		
Low 5	44	8		
High 5	7	97		
Low 6		64	6	
High 6		27	18	1
Low 7		1	87	15
High 7			74	23
Low 8			15	82
High 8				79
Total	200	200	200	200
Mean Grade	4.77	6.12	7.54	8.41
Standard deviation (in terms of year-grades)	.44	.42	.43	.43
Average grade progress per annum over preceding interval		.90	.90	.87
Number of rooms or class-sections represented	32	36	33	29

ing dates just over 80 per cent of the number were located in the same two half-grades. Thus, in 1922, 149 were in Low and High 4; a year and a half later 161 were in High 5 and Low 6; after another year and a half 161 were in Low and High 7; and a year later the same number were in Low and High 8. The progress of the group as a whole, however, is somewhat less regular than this. The figures for grade progress at the bottom of the table show that the 200 average only nine-tenths of a grade a year. In other words approximately 10 per cent have failed of promotion each semester.

This points to a certain selective factor here. The original survey did not include pupils below the Low 4 grade, and the survey four years later did not go beyond High 8. A pupil originally in High 4 who gained even a half-grade extra promotion in the four years would have passed out of the eighth grade before our final test; while laggards originally in fifth grade who failed of promotion at some point would remain in Grade 8 to be measured. Our constant group, therefore, contains somewhat fewer of the very bright and somewhat more than its fair share of pupils who have failed of promotion at some point between Grades 4 and 9.

Despite this fact, it will be seen from Tables III and IV [1] that the group affords a fairly symmetrical distribution of ability, as indicated by quotients and brightness scores in the several subjects. The curves show little tendency to skew toward the lower scores and, while somewhat flat-topped, are probably no more so than would be the case for a single year-grade or age group for New York City schools in general. The mean quotients and brightness scores are close to the 100 and 50 which, respectively, indicate average standing in the three schools as a whole. It appears from these figures that the constant group stands about average in geography and in intelligence, as measured by National Intelligence Test, Scale A, relatively high in reading and arithmetic computation, and low in spelling and arithmetic problems (Otis).

In evaluating the coefficients of correlation to be reported, a

[1] The distributions shown in these tables are in each instance those of the first year in which any one form of the test was given for a second time. For all except the geography and Otis, this year was 1924. The 1924 testing was also the one nearest midway of the four years covered, and fell at a time when the average age of the group was almost exactly 12.0 years (see Table V).

77294

knowledge of the variability of the group is important. As is well known,[2] the size of the coefficient obtained will increase with increase in the variability or dispersion of the group concerned. To permit of comparisons, it is increasingly the practice to report reliabilities as for a group of the variability either of a single age group, such as unselected twelve-year-olds, or of a

TABLE III

QUOTIENTS OF THE TWO HUNDRED PUPILS, INDICATING DISTRIBUTION OF ABILITY IN THE CONSTANT GROUP

Quotient (IQ or EQ)	N.I.T.-A (1924)	Spelling (1924)	Reading (1924)	Geography (1926)	Arithmetic (1924)	Otis (1925)
60– 64	..	..	..	..	..	2
65– 69	..	2	1	..	1	4
70– 74	2	8	2	5	2	9
75– 79	4	10	5	6	7	14
80– 84	15	27	9	15	11	16
85– 89	24	18	19	20	18	22
90– 94	31	34	19	28	20	36
95– 99	38	33	24	15	24	25
100–104	28	27	22	25	30	17
105–109	21	19	25	24	17	17
110–114	21	12	26	17	24	8
115–119	7	4	17	16	12	14
120–124	5	4	12	14	16	5
125–129	2	2	5	5	6	3
130–134	2	..	6	4	3	7
135–139	..	..	4	6	4	1
140–144	..	..	..	..	4	..
145–149	..	..	2	..	1	..
150–154	..	..	1	..	..	..
155–159	..	..	1	..	..	..
Total	200	200	200	200	200	200
Mean	99.1	95.2	105.3	102.6	104.2	96.6
S. D.	11.5	12.0	15.8	15.4	15.7	15.7

[2] Garrett, H. E., *Statistics in Psychology and Education*, pp. 258-60, 271-72. Longmans, Green, 1926.

TABLE IV

BRIGHTNESS SCORES OF THE TWO HUNDRED PUPILS, AS IN TABLE III

Brightness Score	N.I.T.-A (1924)	Spelling (1924)	Reading (1924)	Geography (1926)	Arithmetic (1924)	Otis (1925)
24–26	..	..	..	..	1	..
27–29	1	1	..	..	1	3
30–32	1	2	1	4	1	3
33–35	1	9	3	5	8	9
36–38	13	7	6	6	6	14
39–41	16	25	10	17	15	15
42–44	20	21	16	22	16	26
45–47	32	36	23	28	20	35
48–50	36	43	27	24	20	30
51–53	31	27	29	31	29	22
54–56	24	15	35	18	23	14
57–59	13	9	21	23	23	15
60–62	8	3	13	11	21	5
63–65	3	2	8	9	7	9
66–68	1	..	4	2	7	..
69–71	..	..	3	..	2	..
72–74	..	..	1	..	..	..
Total	200	200	200	200	200	200
Mean	49.4	47.4	52.2	50.1	51.5	47.8
S. D.	6.76	6.61	7.79	7.96	8.90	8.03
S. D. for 12-year-olds in these schools	10.0	*c.* 9.0	10.0	10.0	10.0	10.0

single year-grade, as the sixth or ninth grade. It is interesting, therefore, to note from the last two lines of Table IV that the brightness scores of our constant group on the several tests are only from 68 per cent to 89 per cent as variable as those of an unselected group of twelve-year-olds from the three schools concerned.

Comparison with the variability of a single grade population is made in Table V. Our study of the grade allocation of the constant group (Table II) showed that more than 80 per cent

of the number were in the High 5 and Low 6 grades in May, 1924. But Table V shows that the combined population of these two grades in the twelve New York schools participating in the survey has substantially larger standard deviation than has our constant group.

TABLE V

STANDARD DEVIATION OF AGES AND TEST SCORES OF THE CONSTANT GROUP IN MAY, 1924, AS CONTRASTED WITH THOSE FOR PUPILS OF THE NEAREST CORRESPONDING YEAR-GRADE IN TWELVE NEW YORK CITY SCHOOLS COMBINED

Group	Chronological Age in Months	N.I.T.–A	Spelling (Words Correct)	Reading T Score	Geography T Score	Arithmetic (Examples Correct)
Constant group ...	10.4	18.0	7.83	5.03	8.02	3.75
Grades 5B and 6A combined	15.7	23.1	9.18	8.90	8.88	4.27
Ratio of constant group to year-grade	66%	78%	85%	57%	90%	88%

It is clear that, although our constant group is not confined to a single age or grade group, it is in reality much less variable than would be either of the latter for New York City schools as a whole. It follows that correlations obtained in the present study will be lower in all cases than if based on the total population of a single age or year-grade in the schools concerned.

In spite of this relative homogeneity of the group as regards ability, their training and experience during the four years of the study were highly diversified. The last line of Table II shows that the 200 pupils were distributed among from 29 to 36 different rooms or classes on the several testing dates. Assuming that they were equally scattered in the five semesters when no tests were given, it follows that in the course of the nine semesters of this study, the teachers of 293 different class-sections contributed to the training of our 200 children. So great a diversity of experience might be expected to produce

different rates of progress for different individuals, and so tend further to lower correlations over periods of time, as contrasted with a class of pupils proceeding together through a single school.

DESCRIPTION OF TESTS USED

Tests of six different varieties were used:

1. *Intelligence.* National Intelligence Test: Scale A. World Book Co.
2. *Spelling.* Morrison-McCall Spelling Test. Teachers College, Columbia University, Bureau of Publications. Each scale of this test consists of a selection of fifty words from the Ayres-Buckingham list, dictated with illustrative sentences.

Nifenecker Spelling Test: Form D. Bureau of Reference, Research and Statistics, New York City Board of Education. This test is identical in form with the above, but contains a more difficult selection of words.

3. *Reading comprehension.* Thorndike-McCall Reading Scale. Teachers College, Columbia University, Bureau of Publications. Paragraphs are followed by questions over content, calling for brief answers.
4. *Geography.* New York Standard Geography Test: Series A. Bureau of Reference, Research and Statistics, New York City Board of Education. Each scale consists of a set of twenty brief-answer questions, such as "What state is noted for its manufacture of boots and shoes?" and "What railroad would you take to go from New Orleans to San Francisco?" The different forms, although parallel in scoring, deal with entirely different aspects of geography.
5. *Arithmetic computation.* Woody-McCall Mixed Fundamentals. Teachers College, Columbia University, Bureau of Publications. Each test includes some thirty-five examples of fundamental operations.
6. *Arithmetic problems.* Otis Arithmetic Reasoning Test. World Book Co. Twenty problems of graduated difficulty.

The tests and forms employed on the successive testing dates were as shown on page 18.

In order to permit of rescaling the tests on the basis of results secured from Grades 3 to 8 of the New York Schools, no test form could be used for the two hundred survivors in 1926 which had not been given in the surveys of 1924 or 1925. Accordingly, the 1924 form was used for certain of the tests, and the 1925 form for others. In the case of geography, both forms were given, and the results averaged for greater reliability.

As the Otis test of arithmetic reasoning was unfortunately omitted from the 1924 survey, it is impracticable to average correlations for this test with those for the other five. These correlations serve, however, as a slight, independent check upon conclusions derived from the five tests given throughout.

TESTS USED IN DIFFERENT YEARS

Year	Intelligence	Spelling	Reading	Geography	Arithmetic Computation	Arithmetic Reasoning
1922	National-A Form 1	Nifenecker Form D	Thorndike-McCall Form 1	New York Tests 4 and 5	Woody-McCall Form 2	Otis Form A
1924	Form 1	Form D	Form 1	Test 6	Form 2	
1925	Form 2	Morrison-McCall Form 3	Form 2	Test 7	Form 3	Form A
1926	Form 1	Nifenecker Form D	Form 2	Tests 6 and 7	Form 3	Form A

METHOD OF SCALING TESTS

In the present study it is desired to compare the results obtained with four outstanding types of scores, as follows:

1. Raw scores = number of test items or "points" correct; e.g., on a spelling test the raw score will be simply the number of words spelled correctly.

2. $$\text{T scores} = 50 + 10\left(\frac{\text{Obtained score} - \text{mean for 12-year-olds}}{\text{Standard deviation for 12-year-olds}}\right).$$

T scores were selected as representing the most widely used of current systems of equalizing scale scores in terms of variability, or sigma units. The method of scaling, which assumes a normal distribution of ability in unselected twelve-year-olds, is described in detail in McCall's *How to Experiment in Education,* pp. 98 to 102.

3. Brightness, or B scores = T score + (50 — mean T score for group of pupil's own age).

These scores serve the same ends in a system of variability units as do quotients in an age-scale system. Given equal variability in successive age groups, the above formula would yield

scores identical with a T score based on a group of the pupil's own chronological age, and constancy of sigma status within one's own age group is a logical assumption. (For further description, see McCall, *op. cit.*, pp. 102-9.)

4. Age scores, expressing pupil's performance in terms of months of mental, educational, or subject age.

5. Quotients = 100 MA ÷ CA.

While the significance of such quotients varies with the particular test employed, their practical convenience has led to widespread use.

In order to permit of the above comparisons, it was felt advisable to rescale tests employed in the present study on the basis of results at the disposal of the writer, rather than to accept published norms derived by the various authors from widely disparate groups. This procedure insured:

1. That T scales, age norms, B scores, and quotients for any given test should be derived from the same data, and so be of comparable reliability.

2. That age norms, being based on the average achievement of pupils in the schools from which our group is drawn, should be clearly applicable to our constant group.

3. That validity and reliability achieved should be indicative of that to be expected in the case of tests developed in a given city, or scaled on the basis of results from a periodical survey of its schools.

T scales, B scales, and age scales for the present study were based primarily upon the performance of all pupils in Grades 3 to 8 inclusive, of the three New York City schools described above. For this purpose the results of the 1925 survey were chosen as the most recent and complete. As the tests had not been given to the third grades of these schools, however, and as the inclusion of these grades was indispensable to a satisfactory scaling, the writer obtained permission to apply the tests used in 1925 to all third grades in these schools on the anniversary of the 1925 testing, namely the third week of December, 1926. The returns secured were reduced by a process of random sampling to correspond to the number of pupils in regular daily attendance in the third grade classes of each school as of December, 1925, and the results combined with those already in

hand for Grades 4 to 8. The six-grade composite yielded nearly four thousand pupils, distributed by ages as follows:

Age at Last Birthday	Number of Pupils
7	36
8	377
9	469
10	580
11	616
12	658
13	571
14	390
15	196
16	50
17	6
Total	3949

T scales were developed for each test on the basis of returns from the 658 twelve-year-olds shown,[3] and B corrections found for other ages from eight to fourteen. The procedure followed was that of McCall [4] throughout. B corrections obtained in the case of National Intelligence Test, Scale A, Form 2, are typical of other results, and ran as follows:

Age at Last Birthday	Correction
8	+ 23.9
9	+ 16.9
10	+ 10.1
11	+ 4.5
12	0
13	− 4.4
14	− 7.7

[3] The method of scaling devised by McCall builds on the assumption of a normal distribution of ability in large groups of unselected twelve-year-olds. Study of the city records disclosed that approximately 3 per cent of twelve-year-olds have completed the eighth grade. The selection here would doubtless come chiefly from the ablest 10 per cent of the group. However, this proportion is closely matched by the number of twelve-year-olds of such limited intelligence as to require their segregation in special classes not included in the present surveys. Inquiry of the principals and scrutiny of age and grade distributions failed to disclose further selective factors of importance. The twelve-year-old group above represents, therefore, a distribution of ability presumably symmetrical and approximating the normal, save for a slight truncation at the extremes. As this last characteristic is common to all groups upon which tests are likely to be standardized, no attempt was made to correct for it.

[4] McCall, W.A. *How to Experiment in Education*, pp. 102-09. Macmillan, 1923.

By subtracting each of the above from 50, having regard for signs, the writer arrived at scores which may be taken as closely approximating the mean T scores which would actually be secured from unselected groups of the respective ages, drawn from districts served by the three schools concerned.

The means thus approximated for ages 8 to 14 inclusive in the case of the National Intelligence Test, Scale A, are shown in Figure 1, following, plotted in the form of small circles. By

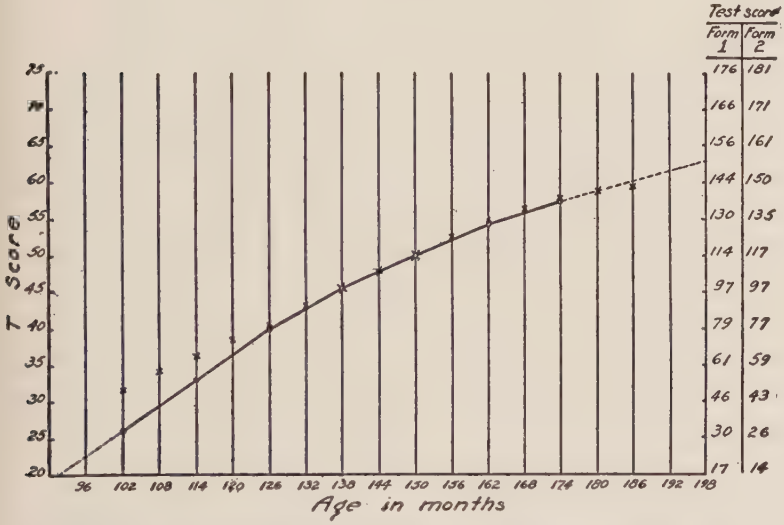


FIGURE 1. Mental Age Equivalents of T Scores by Present Scaling Contrasted with Published Norms of the National Intelligence Test Manual

Circles indicate mean T scores for successive age groups as determined by the writer or, conversely, the mental age equivalents of successive T scores. Figures at right-hand margin show raw test scores on the two forms found to correspond to these T scores. Crosses mark the mental age equivalents of raw scores on Form 1 as shown in test *Manual*, plus 5 points allowance for difference in New York twelve-year-old mean. Divergence of the crosses at lower age levels is attributable to the failure of *Manual* norms to allow for selection in the age groups from which these means were derived.

extending the resulting curve at each extreme, a line is secured from which it is possible to read (1) the B correction suitable for any given month of chronological age, and (2) the T score and mental or educational age equivalent for any given raw score (see figures at right-hand margin of the graph). Tables

giving these equivalents in detail were prepared for all of the tests used.[5]

METHOD OF EQUATING FORMS

Having now a complete set of scales for the test forms used in 1925, there remained the problem of scaling the 1924 forms to match. The ideal procedure in such cases would be to give the several forms to the same four thousand pupils of the three schools concerned, rotating the order for different classes, and scale all by the methods described. In practice, however, the giving of so many tests for the sole purpose of scaling entails an expenditure of time and money quite impracticable in most situations. The device adopted in the present instance is one which should be applicable in the case of other recurrent surveys in large school systems. It may be described as one of equating by percentiles for equivalent groups.

On file in the Bureau of Reference, Research, and Statistics were the usual complete distributions by grades of the scores of pupils in all twelve schools on the successive surveys. These distributions were made with the help of card-sorting and tabulating machines, and were so arranged as to place the survey results in the form most serviceable for general purposes. Merely to compare the returns from a given grade in 1924 with those in 1925 would, of course, be unwarranted if only the latter had enjoyed the advantages of practice effects from a previous survey. Examination disclosed four grades which had experienced the same amount of practice prior to each survey. The High 4 and Low 5 grades of both 1924 and 1925 had participated in no previous survey, while the High 6 and Low 6 had each taken part in the one immediately preceding. In order to secure a fair

[5] A slight departure from the above procedure was made in the case of the geography and spelling tests. As the original scaling of the four geography forms used was effected by Director E. A. Nifenecker by a procedure resembling the above and on the basis of returns from New York schools closely similar to these three, it was found unnecessary to do more than make readjustment of B score and age norms to conform to the standards actually prevailing in the three schools.

In the case of spelling, through the courtesy of Professor McCall the writer was permitted the use of returns obtained by Dr. Morrison from Form 3 of the Morrison-McCall scale given to 13,498 children, constituting the total population of Grades 2 to 8 inclusive in a large number of rural and village schools of New York State. The more adequate sampling of the lower age groups thus rendered available made it desirable to base T scales and B corrections upon this group. The T scores resulting were merely lowered 4.1 points throughout, to make the median performance of twelve-year-olds in the rural schools coincide with that for twelve-year-olds in the three city schools, with their higher level of achievement.

proportion of pupils capable of attempting the more difficult portions of the tests, it was decided to include also the Low 8 grade.[6]

The belief that the performance of pupils in these particular grades should be substantially identical on tests given, as these were in the same week of the term in the two successive years was strengthened by interviews with principals concerned, which indicated that no change in school organization or clientele had occurred in the interval likely to produce a change in standards, and by the fact that average scores for like grades in the two years showed no regular or significant differences.

The number of returns from each grade in 1925 was then reduced to match the number from that grade in 1924 by a process of mathematical sampling from the distribution. There remained for both 1924 and 1925 a composite group of more than six thousand children of equivalent grade status and experience as follows:

Grade	Number
High 4	1,215
Low 5	1,350
Low 6	1,325
High 6	1,332
Low 8	818
Total	6,040

The number of pupils in this group exceeding each given raw score, plus one-half the number getting that score and no more, was calculated for each test form and converted into per cent, after the fashion used in the finding of T scores. It was then a simple matter by equating percentiles to determine the raw test score on the 1925 form which represented an achievement equal to any particular test score on the 1924 form. Thus in Table VI, since 99.1 per cent scored 3 correct answers on Form 1 and 4 correct answers on Form 2, a raw score of 3 on Form 1 is shown in column 4 to be equivalent to a score of 4.0 on Form 2. Then, since a raw score of 4.0 on Form 2 is known (column 5)

[6] Pupils of this grade in 1925 had taken part in two previous surveys, as against one such experience for those in 1924, but it was felt that with their degree of maturity the added practice effect from one additional survey dating three full years previous would be small. In any case the members of this grade constituted less than one-seventh of the final composite.

TABLE VI

EXTRACT FROM TABLES EMPLOYED IN DETERMINING T SCORES FOR THORNDIKE-MCCALL READING FORM 1 BY PERCENTILE EQUATION WITH FORM 2, ALREADY T SCALED

Raw Score	Per Cent Exceeding Plus One-Half Those Reaching This Score		Raw Score on Form 2 Equivalent to Score Shown in First Column on Form 1	T Score on Form 2, as Scaled on Basis of 1925 Results	T Score on Form 1, as Here Found by Interpolation
	On Form 1 (1924)	On Form 2 (1925)			
0	99.9	99.9	0.0	22.6	22.6
1	99.7	99.8	1.5	23.7	24.2
2	99.5	99.6	2.5	24.8	25.2
3	99.1	99.4	4.0	25.6	26.5
4	98.7	99.1	4.6	26.5	26.9
5	98.4	98.4	5.0	28.2	28.2
6	98.1	97.7	5.4	29.1	28.6
7	97.7	96.8	6.0	29.9	29.1
8	97.2	95.6	6.6	30.8	29.6
9	95.8	94.3	7.8	32.0	30.6
10	94.3	92.7	9.0	33.1	32.0
etc.	etc.	etc.	etc.	etc.	etc.

to have a T score value of 26.5, this may be written in the last column as the T equivalent of score 3 on Form 1. Other results are obtained by interpolation.

The method of equating scores here employed possesses the practical advantage of avoiding the necessity for giving tests to an unduly wide range of grades for the purpose of getting adequate age groups. Given a single form of the test, suitably scaled on the basis of age groups, other forms can be calibrated against this at small additional cost, through the medium of score distributions by grades, as tabulated for information of the school authorities. Provided proper precautions are taken to eliminate effects of practice, differences in time of year, and the like, this device has the merit of securing larger numbers than could be obtained in the case of any single age group without great labor and expense.

VALIDITY OF THE SCALING

The method of scaling described as employed in this study is admittedly crude in the extreme. Simplicity and convenience have been favored at the expense of precision. In the T scaling, for example, the procedure originally devised by McCall and since used by other scale makers has been followed without modification. At least two methods alternative to this suggest themselves, either of which might have been employed in the interest of greater refinement. The first would derive three separate T scales based on successive age groups, as eleven-year-olds, twelve-year-olds, and thirteen-year-olds. The three T score values thus determined for each given raw score would then be averaged in order to arrive at the T score equivalent finally adopted.

The second alternative is a modification of the technique developed by Thorndike in transforming original scores on various standard tests into terms of equal units.[7] By this method the sigma or "standard score" equivalents of successive raw scores would first be determined for several different grade populations, taken separately, using the Kelley-Wood Tables of the Normal Probability Integral.[8] The difference between the sigma values of each two successive scores would be computed for each of the grade populations used, and each of these differences then would be multiplied by a fraction equal to the sigma equivalent of a fairly large score interval as derived from a distribution of unselected twelve-year-olds divided by the sigma equivalent of that same interval for the grade concerned. Averaging these transmuted values from the several grades should yield a reliable sigma value for the respective intervals between successive raw scores in terms of the mean-square-deviation of twelve-year-olds. There remains only to multiply these sigma values by 10 and place the 50-percentile performance of twelve-year-olds at 50 in order to arrive at a table of T scores similar to McCall's but much more reliably determined. The determination of B corrections and age norms could continue unchanged.

The writer regrets that the feasibility of the last-named procedure did not occur to him until too late for adoption in the

[7] Thorndike, E.L. *The Measurement of Intelligence*, Chap. VII. Teachers College, Columbia University, Bureau of Publications, 1927.

[8] Kelley, T.L. *Statistical Method*, pp. 373-85. Macmillan, 1923.

present study. Even with the present crude methods, however, two checks available in the case of the National Intelligence Test serve to indicate the substantial validity of the present scaling:

1. By the process of averaging variability scores for different grade groups referred to above, Thorndike[9] has rescaled the National Intelligence Test, Scale A, from scores 20 to 170 inclusive, in units equal in all parts of the range. As a rough check on the like equality of T score units, based on the 658 twelve-year-olds of the present study, this interval from test score 20 to 170 is shown divided into five parts, as follows:

Interval in Original Scores on National Intelligence Test A	Size of Interval in Original Scores	Size of Interval in Scores as Equalized by Thorndike	Size of Interval in T Score Units of Present Scaling (Multiplied by 3 to Facilitate Comparisons)		
			Form 1	Form 2	Mean
20– 50	30	27.6	29.1	29.1	29.1
50– 80	30	24.6	26.4	24.6	25.5
80–110	30	27.6	26.4	22.8	24.6
110–140	30	29.0	30.0	26.1	28.1
140–170	30	39.0	39.6	39.6	39.6

The correspondence of interval values by the present crude scaling with those obtained by Thorndike, while far from ideal, is encouraging. Departures from the original scores are throughout in the same direction as Thorndike's, and the resemblance is closer in the case of Form 1, which has doubtless contributed the greater part of the returns on which Thorndike's calculations were based.

2. In Figure 1 the mental age equivalents of test scores on National Intelligence Test A, as given in Table 3, page 37 of the Fourth Revision of the *Manual,* are indicated by the succession of small crosses. The norms as published have been plotted directly as on the Form 1 scale, save that the test score equivalent to any given mental age has been increased by five points throughout, in order that the norm for 12 years 6 months may fall on score 114, as with our New York group.

[9] *Op. cit.,* p. 244.

The results are almost identical with those of the present scaling for ages from eleven and one-half years (138 months) to fifteen (180 months). Below eleven (130 months) the curves diverge until, by the manual norms, we find the mental age of 8 years 6 months (102) indicated by a test score of 46 (here raised to 51), whereas by present scaling the norm for that age is a T score of 26.1 = test score 34. The reason for this divergence is not far to seek, as the test manual explains that "the norms for the lower ages, especially 8 and 9, are higher than they would be if there were included the duller pupils who have not yet reached the third grade." How highly selected were the eight-year-olds on which the manual norms are based may be judged from the fact that only 2,116 children were found in this age group as contrasted with 5,330 twelve-year-olds. (*Manual,* p. 34.) While recognizing and calling attention to this distortion, the test authors have made no attempt to correct for it in their table of mental age equivalents.

The effect of such overestimation of eight- to ten-year-old performance is to lower unduly the quotients for younger children. By way of indicating the influence of this on constancy of IQ in a group such as our two hundred, the mean test scores for this constant group on successive surveys were converted into IQ's by use of the MA equivalents as published. The results obtained were as shown in the fourth column of Table VII on page 28.

According to the *Manual* norms the group rated six points lower in IQ in 1922 grade than in 1925, whereas the same mode of calculating, using the present New York norms, shows no change. But in 1922 the average chronological age of the group was already ten years and six months. Were measurements available for this same group in 1921, when in the third grade, there is every reason to suppose that quotients would reveal a much more serious lowering at that stage. The last column shows the gradual rise of two points in average IQ's as actually obtained from the present scaling. This increase is too small for statistical significance, and only such as would be expected to result from growing familiarity with tests ways, plus perhaps certain special effort in 1926 (see page 11).

The advantage in the use of local norms appropriate to the group seems evident.

TABLE VII

CONSTANCY OF INTELLIGENCE QUOTIENT IN CONSTANT GROUP WHEN BASED ON PRESENT SCALING AND LOCAL NORMS AS CONTRASTED WITH THOSE OF THE NATIONAL INTELLIGENCE TEST MANUAL

Year of Testing	Mean Grade Location of Group	Mean Score on N.I.T.–A	IQ Equivalent to Mean Test Score Obtained		Mean of Individual IQ's as Actually Computed by the Norms of the Present Scaling
			As Based on Norms in Test Manual	As Based on Norms of Present Scaling	
1922	4.77	68.0	94	96	98 ± 0.6
1924	6.12	98.8	99	97	99 ± 0.5
1925	7.54	122.4	100	96	98 ± 0.6
1926	8.41	131.7	99	96	100 ± 0.7

THE CRITERION OF MEASUREMENT

The criterion of reliability employed in this study is the success with which any measure or combination of measures is able to forecast the status of members of the constant group in December, 1926. As against this final or criterion measurement, the reliability of predictions obtained from the December, 1925, scores alone is compared with that obtained from a combination of the 1925 with 1924 results, of the 1925 with 1922 results, and of all three together.

The accuracy of the correspondence between predicted and measured status is indicated by three distinct criteria:

1. The size of the obtained coefficient of correlation with 1926 results.

2. The size of the radical $\sqrt{1-r_{11}}$, or the ratio of the standard error of measurement to the standard deviation of the scores.[10] The use of the standard error of measurement for this purpose, rather than the standard error of estimate, has the advantage of correcting for the unreliability which is due to inadequacy of the criterion. The size of this measure thus shows the degree of correspondence, or rather the lack of correspondence, with

[10] Franzen, Raymond. "Statistical Issues." *Journal of Educational Psychology*, XV: 378-82, Sept., 1924.

the probable true ability of the individual as this would be obtained from the average of scores on an infinite number of forms of the test, instead of the single form actually applied in December, 1926.

3. The mean differences between quotients and brightness scores obtained in 1926 and those obtained from previous measurements or simple averages of these measurements. These differences serve to indicate the degree of correspondence actually secured in practice under conditions of the present study, when no attempt is made to correct for changes in means and standard deviations of scores in successive years, due to practice effects and the like.

CHAPTER III

RESULTS: THE INFLUENCE OF TIME UPON SCORES

The record of each of the two hundred pupils who had taken all six tests on all four occasions was transferred to a special individual record card, as illustrated on the following page. The lines on these cards prefaced by the numbers 22, 24, 25, and 26, show the performance of the pupil in the calendar years of 1922, 1924, 1925, and 1926, respectively. In the interests of accuracy the raw scores made by each pupil were taken directly from the teacher's class record sheets, as compiled at the time of the different surveys. These raw scores were entered in the second columns of the cards under the heading of "No." signifying "Number of test points correct." From these raw scores, by means of the scales derived as described in Chapter II, the corresponding T scores, B scores, mental ages, and quotients were computed and entered in the columns designated. In the three spaces of the B score column lying immediately below the B score on each test for the year 1925, were then entered, respectively, the mean B score for the years 1925 and 1922, the mean for 1925 and 1924, and the mean for 1925, 1924, and 1922 combined. Corresponding entries in the Q columns show the mean quotients for the same combinations of years.

In the section of the card headed "NIT" and bearing the results from the National Intelligence Test, the space immediately below the T score for 1925 contains the T score which the pupil must have made in 1926 in order to obtain the same B score he secured in 1922. The space next below that shows the T score as predicted for 1925 on the like assumption of constancy on the 1922 B score. Below that and next above the T score for 1926, is shown the T score required for 1926 in order to maintain the same B score as in 1924. The corresponding entries in the Mental Age column give, respectively, (1) the MA predicted for 1926, assuming the IQ of 1922 to remain constant,

(2) the MA for 1925 similarly predicted, and (3) the MA for 1926 as predicted by the IQ of 1924.

In reckoning the chronological ages entered at the top of the cards, and in calculating the IQ's and B scores as above, the 1924 tests were accounted as having been taken at a time mid-way between the 1922 and 1925 surveys. This is in accordance with their position in reference to school years, rather than strict calendar dates.

NAME P, John A. B OR G. SCHOOL 165 BORN 3-21-12

CLASS 5A² 6B² 7B' 8B' AGE 129 147 165 177

NIT	No.	T	B	M.A.	Q
22	111	49	58	148	115
24	110	49	50	147	100
25	155	62	57	193	117
		66	57.5	203	116
		67	53.5	189	108.5
		58	55	177	111
26	162	68	60	222	125
READ					
22	21	49	41	148	115
24	25	57	58	175	119
25	27	65	61	203	123
			51		119
			59.5		121
			53		119
26	29	67	59	210	119
SPELL					
22	5	31	38	102	79
24	18	41	42	126	86
25	36	48	43	144	87
			40.5		83
			42.5		86.5
			41		84
26	28	47	37	142	80
GEOG					
22	7.5	48.5	61	153	119
24	9	56	60	178	121
25	17	71	69	237	144
			65		131.5
			64.5		132.5
			63		128
26	14	66.5	60	219	124

OTIS	No.	T	B	M.A.	Q
22	10	53	61	161	125
24			(59)		(120)
25	13	61	57	189	115
			59		120
26	13	61	54	189	107
ARITH.					
22	21	39	49	127	98
24	28	51	52	153	104
25	25	47	42	144	87
			45.5		92.5
			47		95.5
			48		96
26	27	51	43	153	86
Educ					
22		44	60	138	107
24			54		110
25		58	54	183	111
			52		109
		54	54	177	110.5
		57	53	189	109
26		58	51	183	103

FIGURE 2. Sample Card Used for the Compilation of Individual Records

In the seventh section of the card, under the heading "Educ," are shown the mean T scores, B scores, and quotients made on the five educational, or achievement, tests in the respective years. These means were obtained by a simple averaging of the corresponding scores as shown under the respective tests, exclusive of the National Intelligence. To permit of including Otis test results in arriving at this rough average of educational, as contrasted with intelligence test, status, each pupil was reckoned as having earned a B score and Q on the Otis in 1924 equal to the mean of his 1922 and 1925 performances. Once calculated, these "Education" scores were treated exactly as though obtained in

the usual fashion from an achievement test battery of duly weighted components.

A complete record of the original data for each of the pupils of the constant group is on file in Teachers College Library, Columbia University, for the use of those interested.

INTERCORRELATIONS OBTAINED

The intercorrelations among the results of the four testings, as calculated by the Pearson product-moment formula, are shown for each test and type of score in Tables VIII to XI. The approximate linearity of the regressions was confirmed by inspection in the case of all tests except that of spelling for 1926. In this one instance the Nifenecker Form D used proved either too easy or too familiar for the group. Sixteen per cent of the pupils missed less than two of the fifty words. The result was a distortion of the line of relationship which manifests itself in a lowering of the correlations involving 1926.

TABLE VIII

COEFFICIENTS OF CORRELATION FOR RAW SCORES ON TESTS

Years Correlated	Number of Years Interval	Test							Mean *r* of the Five Tests Other Than Otis	P.E. for a Single *r* of This Size
		N.I.T.–A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean T Score on Educational Tests*		
1926, 1925 ..	1.0	.807	.736	.614	.692	.618	.695	.789	.693	±.025
1926, 1924 ..	2.5	.761	.734	.556	.641	.598	...	...	.658	±.027
1926, 1922 ..	4.0	.631	.665	.515	.382	.442	.537	.645	.527	±.034
1925, 1924 ..	1.5	.747	.751	.528	.586	.570	...	...	.636	±.028
1925, 1922 ..	3.0	.610	.634	.497	.255	.385	.526	.587	.476	±.037
1924, 1922 ..	1.5	.657	.804	.632	.393	.506	...	...	.598	±.031
Mean	2.25	.702	.721	.557	.491	.598			.598	±.031

* In determining the average standing on the five achievement or educational tests, T scores were used, since the simple averaging of raw scores on the various tests would be impracticable.

TABLE IX

Coefficients of Correlation for T Scores

Years Correlated	Number of Years Interval	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
		N.I.T. –A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean T Score on Educational Tests		
1926, 1925 ..	1.0	.809	.684	.603	.680	.610	.688	.749	.677	±.026
1926, 1924 ..	2.5	.743	.689	.551	.645	.560	...	...	.638	±.028
1926, 1922 ..	4.0	.650	.669	.482	.401	.435	.546	.645	.527	±.034
1925, 1924 ..	1.5	.711	.754	.537	.577	.549	...	...	.626	±.029
1925, 1922 ..	3.0	.601	.653	.534	.273	.385	.498	.587	.489	±.036
1924, 1922 ..	1.5	.651	.785	.654	.365	.524	...	...	.596	±.031
Mean	2.25	.694	.706	.560	.490	.510	...	...	.592	±.031

TABLE X

Coefficients of Correlation for Quotients

Years Correlated	Number of Years Interval	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
		N.I.T. –A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean Quotient on Educational Tests		
1926, 1925 ..	1.0	.833	.782	.644	.754	.644	.758	.853	.731	±.022
1926, 1924 ..	2.5	.823	.782	.594	.731	.575	...	.853	.701	±.024
1926, 1922 ..	4.0	.752	.775	.569	.581	.525	.630	.814	.640	±.028
1925, 1924 ..	1.5	.800	.824	.665	.668	.609	...	.859	.713	±.023
1925, 1922 ..	3.0	.741	.786	.678	.550	.483	.646	.808	.648	±.028
1924, 1922 ..	1.5	.790	.865	.740	.712	.632	...	.898	.748	±.021
Mean	2.25	.790	.802	.648	.666	.578	.678	.848	.697	±.024

TABLE XI

COEFFICIENTS OF CORRELATION FOR BRIGHTNESS SCORES

Years Correlated	Number of Years Interval	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
		N.I.T.–A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean B Score on Educational Tests		
1926, 1925 ..	1.0	.852	.797	.656	.763	.651	.758	.860	.744	±.021
1926, 1924 ..	2.5	.835	.790	.622	.749	.627	...	.859	.725	±.023
1926, 1922 ..	4.0	.750	.763	.579	.613	.554	.660	.811	.652	±.027
1925, 1924 ..	1.5	.820	.851	.663	.709	.624	...	.878	.733	±.022
1925, 1922 ..	3.0	.738	.810	.629	.569	.525	.653	.805	.654	±.027
1924, 1922 ..	1.5	.791	.858	.726	.704	.670	...	.906	.750	±.021
Mean	2.25	.798	.812	.646	.685	.609	.690	.853	.710	±.024

In these tables the column immediately following the Otis arithmetic gives the intercorrelations among the educational battery scores obtained as explained in the preceding section. These should not be confused with the figures of the next to the last column, which consist of the arithmetic means, or simple averages, of the coefficients obtained for the five tests, National Intelligence to Woody-McCall, inclusive.

It will be noted that the relative educational status, as indicated by the mean of quotients as well as B scores on the five achievement tests, shows a constancy appreciably above that for the one form of National Intelligence Test A. This finding throws little light upon the constancy of achievement as contrasted with intelligence, since the achievement tests involved represent a total testing time at least two and one-half times as great as that devoted to the National Intelligence. This greater adequacy of testing would in itself make for increased reliability. There have, however, been many well-founded objections to the computation of age scores, quotients, and the like on the basis of ordinary point scales, and particularly to the use of such

scores as though directly comparable between test and test, regardless of the mode of standardization employed by the different authors.[1] In view of these theoretical shortcomings, it is interesting to note the high self-correlations and relative constancy (cf. Table XXIV) found actually to prevail in the case of an average of such scores derived from a number of tests of uniform standardization.

RELIABILITY OF THE COEFFICIENTS

The final columns of Tables VIII to XI show the probable errors of coefficients of the size of the mean correlations just preceding. *These are not intended to indicate the P.E. of the mean correlations as shown,* since these could not be determined without a knowledge of the intercorrelations among the different types of test.[2] The purpose here is merely to indicate the proximate order of reliability of the coefficients for the individual tests. As $N = 200$ for all correlations of the present study, the reliability of any given coefficient can readily be determined from the formula, $P.E._r = \pm .0477 (1 - r^2)$, or read with reasonable accuracy by simple interpolation in Table XII, following:

TABLE XII

PROBABLE ERRORS FOR COEFFICIENTS OF CORRELATION BASED ON A GROUP OF TWO HUNDRED

r	P.E.	*r*	P.E.	*r*	P.E.
.118	.046	.573	.032	.789	.018
.276	.044	.608	.030	.815	.016
.344	.042	.641	.028	.840	.014
.400	.040	.674	.026	.865	.012
.449	.038	.704	.024	.889	.010
.494	.036	.733	.022	.912	.008
.535	.034	.762	.020	.935	.006

The computation of the reliability of differences between correlations such as the foregoing is somewhat complicated. The

[1] Gates, A.I. "The Unreliability of MA and IQ based on Group Tests of Mental Ability." *Journal of Applied Psychology,* VII: 93-100, 1923.

Miller, W.S. "The Variation and Significance of Intelligence Quotients from Group Tests." *Journal of Educational Psychology,* XV: 359-66, 1924.

Rand, Gertrude. "A Discussion of the Quotient Method of Specifying Test Results." *Journal of Educational Psychology,* XVI: 599-618, 1925.

Freeman, F.N. *Mental Tests,* Chaps. 11 and 13. Houghton Mifflin, 1926.

[2] Yule, G.U. *An Introduction to the Theory of Statistics,* pp. 210-11. London: Chas. Griffin and Co., 1924.

procedure may be illustrated in the case of the difference between $r_{1926,\ 1925}$, and $r_{1926,\ 1922}$ for B scores, as shown in Table XI. Let $r_{1926,\ 1925} = r_{12}$, $r_{1926,\ 1922} = r_{13}$, and $r_{1925,\ 1922} = r_{23}$. Let σ_{12} = Standard Error of r_{12}, and $k_{12} = \sqrt{1 - r^2_{12}}$, etc. Then the Standard Error of the Difference between r_{12} and $r_{13} = \sqrt{\sigma^2_{12} + \sigma^2_{13} - r_{r12\ r13}\ \sigma_{12}\ \sigma_{13}}$ in which [3]

$$r_{r12\ r13} = r_{23} - \frac{r_{12}\ r_{13}\ (k^2_{23} - r^2_{12} - r^2_{13} + 2\ r_{12}\ r_{13}\ r_{23})}{2\ k^2_{12}\ k^2_{13}}$$

Applying the above to the coefficients of Table XI, we obtain:

Test	Difference	Standard Error of Difference	Difference / Standard Error
Intelligence	.102	.0285	3.58
Spelling	.034	.0295	1.15
Reading	.077	.0467	1.65
Geography	.150	.0411	3.66
Woody-McCall	.101	.0480	2.10

While it is evident that these differences, particularly in the case of spelling and reading, fall short of the common statistical criterion of three times the standard error, yet it will be recognized that the mean of five such differences, all of which are positive or of the same direction, will possess a highly satisfactory degree of reliability. As the computations are, manifestly, exceedingly laborious and not particularly rewarding, the writer has not attempted to compute the probable errors of the multitudinous differences entailed in the various tables of this study. The example given may enable the experienced reader to estimate by inspection and comparison the general order of reliability in other cases.

INFLUENCE OF TIME ON CORRELATIONS

Prolonged study of the coefficients in Tables VIII to XI has revealed only one characteristic which can safely be affirmed for all tests and all types of score. This is the tendency for correlations to decrease with increasing time interval.

Even to this generalization exceptions must be made in the case of particular correlations or particular tests. There is, for example, a noticeable tendency for correlations of 1926 results with 1924 or 1922 to equal or exceed the correlations of 1925 with the same years. This is true for practically all types of score

[3] Kelley, T.L. *Statistical Method*, p. 179, formula 129.

and test except the spelling, which, as has been stated, proved somewhat too easy for measuring adequately the abler pupils in 1926. This greater reliability of the 1926 measurement might be explained in the case of geography by the fact that two forms of the test were used in 1926 as against one in 1925. In intelligence, the test form used in 1926 was the same as that in 1924 and 1922, and different from that of 1925. However, the tendency is almost as clearly defined in the case of the Thorndike-McCall, Woody-McCall, and Otis arithmetic tests where 1926 enjoyed no such advantage. This may point merely to a reduction in errors of marking owing to the extraordinary pains taken to insure accuracy in the 1926 results,[4] but there seems at least some ground for supposing that the same factors which caused the greater reliability noted in the later measurements on Baldwin's group[5] have been at work here also. It is not unreasonable to suppose that increasing habituation of the pupils

TABLE XIII

MEAN COEFFICIENTS OF CORRELATION OVER VARYING TIME INTERVALS

1 Type of Score	2 Years Interval Between Tests	3 Mean Interval	4 Number of r's	5 Intelligence	6 Spelling	7 Reading	8 Woody-McCall Arithmetic	9 Mean of the Four r's	10 New York Geography	11 Mean Educational Score
Raw score	1.0–1.5	1.33	3	.737	.764	.591	.565	*.664*	.557	*
	2.5–3.0	2.77	2	.686	.684	.526	.491	*.597*	.448	*
	4.0–	4.00	1	.631	.665	.515	.442	*.563*	.382	*
T score	1.0–1.5	1.33	3	.723	.741	.598	.561	*.656.*	.541	*
	2.5–3.0	2.77	2	.672	.671	.543	.473	*.590*	.459	*
	4.0–	4.00	1	.650	.669	.482	.435	*.559*	.401	*
Quotient	1.0–1.5	1.33	3	.808	.824	.683	.628	*.736*	.711	.870
	2.5–3.0	2.77	2	.782	.784	.636	.529	*.683*	.641	.831
	4.0–	4.00	1	.752	.775	.569	.525	*.655*	.581	.814
B score	1.0–1.5	1.33	3	.821	.835	.682	.648	*.747*	.725	.881
	2.5–3.0	2.77	2	.787	.800	.626	.576	*.697*	.659	.832
	4.0–	4.00	1	.750	.763	.579	.554	*.662*	.613	.811

* Correlations with 1924 were not computed, owing to lack of Otis arithmetic scores.

[4] In addition to the employment of especially competent examiners, all scoring and computation was checked by two or more persons independently.

[5] Page 4.

to the taking of tests as well as growing skill on the part of the examiners will, other things being equal, make for increasing reliability in successive measurements.

To average out such extraneous factors as position in the series, form of test used, and the like, and to secure as direct an indication as possible of the influence of time alone, the three coefficients available in the case of each test and type of score for intervals of one to one and a half years have been averaged; likewise the two for intervals of two and a half and three years. The results, as shown in Table XIII, reveal a progressive and highly consistent decline in correlation with each increase in time interval. This holds true without exception for all tests and types of score for which complete returns are available.

An excellent idea of the general relationship may be obtained from column 9, which gives the mean for the preceding four tests. From this average the results for geography were excluded, since the use of two forms in 1922 and 1926 against one in 1924 and 1925 rendered correlations and standard deviations not strictly comparable. The line of these means, as of column 9, is shown graphically in Figure 3. It will be seen:

1. That the line of relationship for each type of score closely approaches the rectilinear.[6] Certainly the greater part, if not all, of the upturn noted at four years is due to the fact that 100 per cent of the correlations for the 4-year interval involve the more reliable 1926 measurement, as against 50 per cent of those for the 2.77 year, and 33⅓ per cent for the 1.33 year. It may be added that inclusion of geography in the averages would have made the straight-line relationship even more striking.

2. That the effects of time elapsed, as measured by the Pearson coefficient (r), are only slightly greater in the case of raw scores or T scores than in the case of quotients or B scores. When comparison is made in terms of the coefficient of alienation (k), this difference becomes negligible.

3. That there is a slight but fairly consistent tendency for correlations in the case of quotients to fall short of those for B scores, and a similar but less marked inferiority of T scores to raw scores. The difference in favor of the coefficients for B

[6] It is interesting to note that this finding accords admirably with the assumption which Gates was led to make in connection with correlations obtained from a very different type of group over intervals ranging from four to twenty months only. See page 526 of reference 21 in Chapter I.

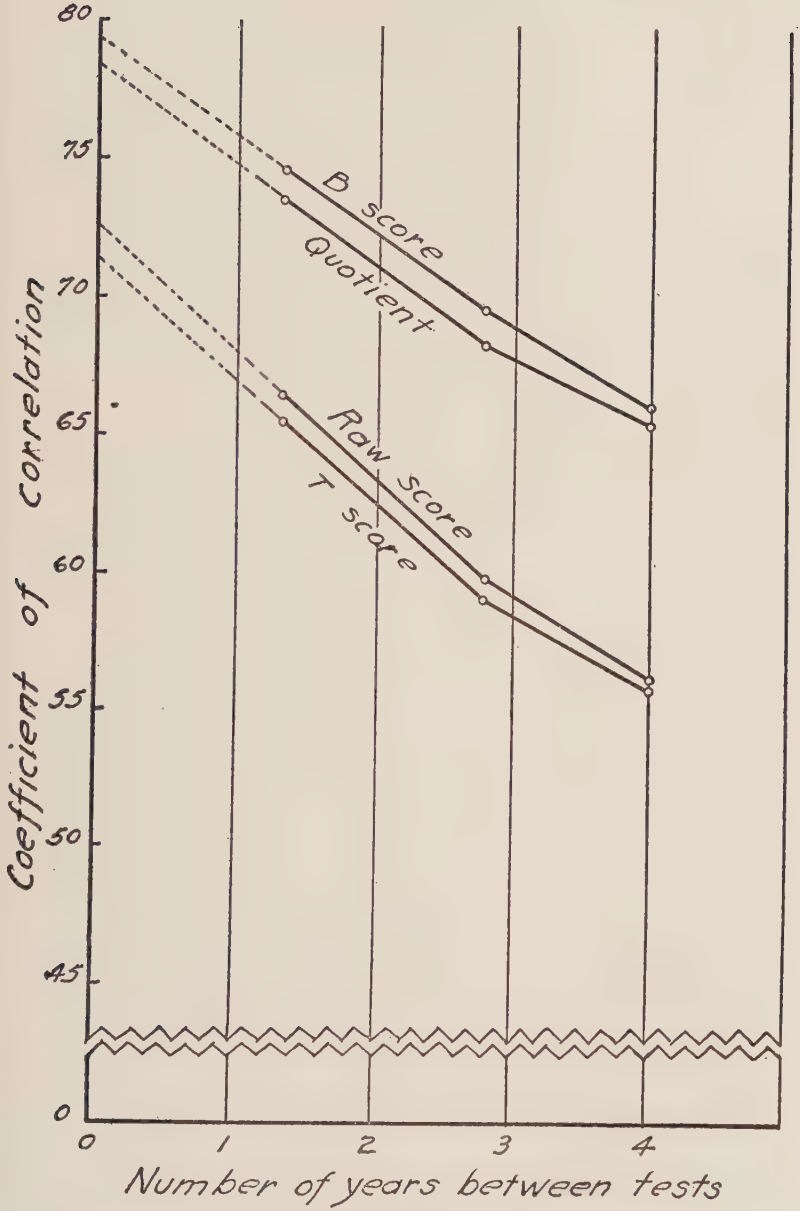


FIGURE 3. Decrease in Correlation with Increase in Time Interval Between Tests

scores as opposed to quotients averages just .013 ± .013.[7] Such a difference of but 1 P.E. would occur about once in four times by chance alone, so that little significance can be attached to it.

On the strength of the evidence for a nearly rectilinear relation, as noted under 1, the lines of Figure 3 are shown extended in the form of dashes as far as the one-year interval, and thereafter to 0 as a dotted line, to indicate the greater uncertainty of that portion.[8] If we accept these extensions as at least approximately correct, it becomes possible to read from the point of intersection with the vertical for 0 years interval, the reliability coefficient which corresponds to the obtained correlations shown for any given number of years interval. The coefficients thus read from the several lines are given in Table XIV.

TABLE XIV

CORRESPONDING COEFFICIENTS OF CORRELATION BETWEEN TESTS OF EQUAL RELIABILITY GIVEN TO LIKE GROUPS AFTER DIFFERENT INTERVALS OF TIME

Type of Score	Years Interval Between Tests				
	0	1	2	3	4
Raw score	.725	.679	.633	.590	.563
T score	.715	.670	.625	.584	.559
Quotient	.783	.747	.711	.678	.655
B score	.791	.757	.723	.690	.662

Letting r_{11} signify the coefficient obtained over a given time interval, and r_{21} the coefficient over a different interval but with

[7] Compare figures at extreme right of the bottom lines of Tables X and XI.

[8] The conflict with Woodyard's findings at this point is more seeming than real. For periods measured in hours, days, and weeks the evidence is clear that the effect of time elapsed is negligible in comparison with other factors making for variability. Her conclusions with reference to periods in excess of four months are, however, based exclusively on the data from two sets of retests over an interval of one year. Of these, one relates to the Thorndike Intelligence Examination given to normal school students, the other to the closely similar I.E.R. tests used in first and second year of high school. Conclusions derived from tests of general intelligence with groups of this degree of maturity are clearly of doubtful application to measurements of school achievement in the case of elementary and junior high school pupils. Woodyard, herself, has been careful to point out that of fifteen tests given seventh grade children, nine show decreased correlation over an interval of four months against four showing increased (cf. page 5).

the same type of score, and letting ρ indicate the ratio of $\sqrt{1 - r_{1I}}$ to $\sqrt{1 - r_{2II}}$, it will be found that the slopes of the lines of Figure 3 are such that this ratio is substantially the same for all four types of score from 0 to 3 years interval, the range throughout which the determination is most reliable (see Table XIII). By averaging the four ρ's obtained we thus arrive at a constant,

$$c = \frac{\sqrt{1 - r_{1I}}}{\sqrt{1 - r_{2II}}}, \qquad (1)$$

applicable to all four types of score. Knowing c, it becomes a simple matter to predict the r_{1I} to be expected from a test yielding r_{2II} over a different time interval. Solving equation (1), we obtain

$$r_{1I} = 1 - c^2 + c^2 r_{2II} \qquad (2)$$

To illustrate, let

r_{1I} = correlation between two forms of test for 0 years interval, or the usual reliability coefficient, and

r_{2II} = correlation between same two forms given 1 year apart.

Reading from Table XIV, we obtain:

	r_{1I}	r_{2II}	ρ
Raw score	.725	.679	.924
T score	.715	.670	.930
Quotient	.783	.747	.926
B score	.791	.757	.927
			c = M = .927
			c^2 = .86

Whence, $r_{1I} = .14 + .86\, r_{2II}$ (3)

To illustrate once more, let

r_{1I} = reliability coefficient, as above, and

r_{2II} = correlation between tests given 1.5 years apart.

Then, reading from the lines of Figure 3, we obtain

	r_{1I}	r_{2II}	ρ
Raw score	.725	.656	.894
T score	.715	.647	.899
Quotient	.783	.729	.895
B score	.791	.740	.897
			c = M = .896
			c^2 = .803

Whence, $r_{1I} = .197 + .803\, r_{2II}$ (4)

Like computations yield for r_{2II} = correlation between tests given 2.0 years apart, the equation,

$$r_{II} = .246 + .754\, r_{2II} \qquad (5)$$

and for r_{2II} = correlation between tests given over a three-year interval, the equation

$$r_{II} = .324 + .676\, r_{2II} \qquad (6)$$

Similar equations may be developed for prediction of the correlation over any given interval up to a limit of three years if we know the correlation obtained over another interval within these same limits; and when applied to the data of Tables VIII to XI, or XV, will be found to yield very satisfactory results.

COEFFICIENTS OF RELIABILITY

By assuming, as seems reasonable, that the relations signified by the above equations may be trusted to hold good also for coefficients somewhat above and below the range from .55 to .80 for which they were derived (Figure 3), we are now in position to determine approximately the reliabilities of the several tests employed in the present study.[9]

The fourth line of Table XV gives the coefficients as obtained between T scores for our constant group over intervals of 1.0 to 1.5 years. In all cases except the Otis and Education batteries, for which no 1924 scores are available, the correlations taken are between 1924 and 1925. For these two years the test forms used were in no case the same, and the testing was carried out under survey conditions by regular school teachers rather than the writer's special examiners.

Kelley has developed a formula for the reduction of an obtained coefficient for a group of known variability to the corresponding reliability for a second group of different but known variability, when the measure itself is equally reliable through both ranges. The equation is

$$R = 1 - \frac{\sigma^2 (1 - r)}{\Sigma^2}$$ (p. 222, formula 178, Kelley: *Statistical Method*)

where σ or r represent the standard deviation and reliability coefficient for the group of smaller range of talent, and Σ and R

[9] Evidently these equations cannot be applied to coefficients too far below the range here indicated, as it would manifestly be unsafe to argue from (3) that tests yielding a correlation of .00 over a one-year interval would show a correlation of .14 on immediate repetition. But no such extreme is here contemplated.

TABLE XV

COEFFICIENTS OF RELIABILITY FOR TESTS OF THE PRESENT STUDY EXPRESSED AS FOR GROUPS OF THE VARIABILITY OF TWELVE-YEAR-OLD CHILDREN OF THE SCHOOLS CONCERNED, AND CORRECTED FOR THE EFFECTS OF TIME ELAPSED BETWEEN TESTS

	National Intelligence Test A	National Intelligence Test A	Nifenecker and Morrison-McCall Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Mixed Fundamentals	Otis Arithmetic Reasoning	Educational Battery (Mean Score on Five Tests)
Years Compared	1924, 1925	1925, 1926	1924, 1925	1924, 1925	1924, 1925	1924, 1925	1925, 1926	1925, 1926
Test Forms used	1, 2	1, 2	3, D	1, 2	6, 7	2, 3	1	
Years Interval	1½	1	1½	1½	1½	1½	1	1
σ of T scores	5.26	6.00	5.49*	6.52	7.57	7.26	7.22	4.76
Correlation (r) as obtained with T scores	.711 ±.024	.809 ±.016	.754 ±.021	.537 ±.034	.577 ±.032	.549 ±.033	.688 ±.025	.749 ±.021
Correlation (r) after correction for effects of time elapsed between tests by use of equations (3) & (4)	.768	.831	.802	.628	.660	.638	.731	.784
Reliability as for group of 12-year-old variability, using the Kelley formula	.920 ±.010	.931 ±.007	.926* ±.009	.803 ±.025	.757 ±.028	.762 ±.027	.837 ±.017	.943 ±.007
Estimated reliability coefficient as for a group of the variability of 12-year-olds, after correction for the effects of time elapsed between tests	.936	.941	.941*	.842	.805	.809	.850	.951

* Since the spelling tests as used were T scaled on the basis of unselected twelve-year-old children in New York state rural and village schools, the corrected reliabilities are as for such a group, and slightly higher than if based on New York City schools alone.

the corresponding measures for the group of wider range. As the correlations of Table XV are in all cases for T scores, and the standard deviation of T scores for all twelve-year-old children in Grades 3 to 8 of our three New York schools is by definition and derivation equal to 10.0 (Chapter II), it is a simple matter to arrive at the estimated reliabilities for twelve-year-olds, as shown in the last two lines of the table. The probable errors of the r's as estimated in the next to the last line have been computed from the formula derived by Shen.[10]

$$\text{P.E.}_{\text{R}} = \frac{.6745\ \sigma_r\ \text{RK}^2}{rk^2}$$

The very satisfactory correspondence achieved in the final results is illustrated in the case of National Intelligence Test A. This shows a coefficient of reliability of .936 for twelve-year-olds as estimated from the 1924-25 coefficient, compared with .941 as estimated from the 1925 and 1926.

For at least three of the tests used, the findings of other investigators are available in a form which renders direct comparison possible. Symonds [11] has reported reliability coefficients for the National Intelligence Test, Scale A, Thorndike-McCall reading, and Woody-McCall arithmetic secured in Honolulu from a composite group of Grades 4 to 8 but with full data as to standard deviations supplied. Willson [12] has converted the correlation for the National Intelligence Test into the equivalent coefficient for unselected twelve-year-olds, and Kelley [13] has shown the variability of a composite group of Grades 4 to 8 in general school achievement to compare with that of unselected twelve-year-olds in the ratio of 54.4 to 57.2. We are thus enabled to render Symonds' data as follows:

Test	Forms	Results Obtained from the Five-Grade Composite			Estimated r for Unselected 12-Year-Olds
		N	r	P.E.	S.D. (months)	
N.I.T.–A	1, 2	232	.922 ± .007	23.9	.934 ± .006
Thorndike-McCall	1, 2	232	.794 ± .016	22.2	.814 ± .015
Woody-McCall	1, 2	229	.855 ± .012	25.7	.869 ± .011

[10] Shen, E. "Standard Error of Certain Estimated Coefficients of Correlation." *Journal of Educational Psychology*, XV: 7, Oct., 1924, Formula 1*b*.

[11] Symonds, P.M. "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Journal of Educational Psychology*, XV: 423-32.

[12] Willson, G.M. "Deviations of Age Scores and Quotients." *Journal of Educational Psychology*, XVI: 203, March, 1925.

[13] Kelley, T.L. *Interpretation of Educational Measurements*, p. 198. 1927.

Miss Mary Cronin, in an unpublished Master's thesis of Stanford University entitled "The Reliability of Certain Educational Tests," reports reliability coefficients for the reading and arithmetic tests as for groups of the variability of one-year grade, as follows:

Thorndike-McCall	Forms 2, 4	$r = .567 \pm .049$
Woody-McCall	Forms 1, 2	$r = .696 \pm .046$

Comparing the last line of Table XV with Symonds' results reveals a most satisfactory correspondence. The maximal difference is, in fact, less than twice its probable error, and the reliabilities of the three tests by the present findings average practically the same as with Symonds'. The same can be said of the agreement between Cronin's reliabilities and those of the sixth line of the table, with which they are properly comparable. The mean reliability in the two instances is, in fact, almost identical. That this should be true for tests given as were these by so great a number of different classroom teachers and in so wide a variety of times and places, is a striking testimony both to the extraordinary care with which the New York surveys have been conducted and to the remarkable possibilities in the widespread use of standard tests. The closeness of the correspondence goes far, also, to confirm the validity of the correction formulas as derived.

CURVES OF IMPROVEMENT

Curves of improvement in the various functions tested are doubtless best shown in terms of T scores, representing a scale of presumably equal units. The mean T scores of the six tests, and the mean educational score on the five tests of achievement at the four times of testing are given in Table XVI and graphically depicted in Figure 4. Dashed lines in this figure connect the 1922 and 1925 means on the Otis and Education scores, for which no 1924 measurement is available.

The course of the several curves agrees fairly closely, with performance in geography, expressed in terms of Nifenecker's norms,[14] lagging somewhat below the rest. The general picture is the usual one of negative acceleration, interrupted by a somewhat unexpected rise in 1926. For this last, three explanations suggest themselves.

[14] See Chapter II, note 5.

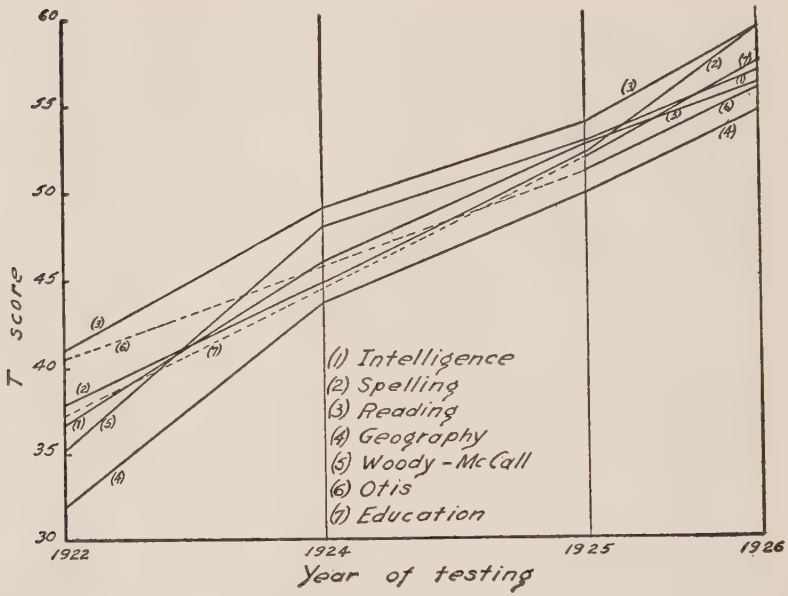


FIGURE 4. Curves of Improvement in the Six Functions and Educational Battery Score, as Measured by Mean T Scores of Successive Testings

TABLE XVI

Curves of Improvement on the Six Tests and Educational Battery Score as Measured by Mean T Scores of the Constant Group at the Four Times of Testing

Test	Dec., 1922	May, 1924	Dec., 1925	Dec., 1926.
Intelligence	36.9 ± 0.3	46.1 ± 0.3	52.6 ± 0.3	56.5 ± 0.3
Spelling	37.0 ± 0.3	44.8 ± 0.2	52.1 ± 0.3	59.2 ± 0.5
Reading	41.0 ± 0.3	49.1 ± 0.3	53.8 ± 0.3	59.0 ± 0.4
Geography	31.9 ± 0.3	43.7 ± 0.4	49.8 ± 0.3	54.5 ± 0.3
Woody-McCall	35.3 ± 0.3	48.2 ± 0.4	52.7 ± 0.3	55.7 ± 0.4
Otis arithmetic	40.7 ± 0.3		50.9 ± 0.3	55.5 ± 0.4
Mean T score (ex. Otis)	36.6	46.4	52.2	57.0
Mean educational score	37.4 ± .14		51.8 ± .14	56.8 ± .18

1. The rise may be in part the reflection of some peculiarity in the present scaling. As against this it should be stated that the angle in the case of Thorndike-McCall reading, for example, is quite as sharp in terms of the author's scaling as by the present writer's.

2. As remarked in Chapter II, the feeling on the part of the children that they had been singled out for the distinction of a final test on grounds creditable to themselves appeared to provide an unusual incentive to effort on the 1926 test.[15]

3. It is probable that the 1925 record may have been somewhat depressed by the fact that for all tests except the Otis, wholly new forms were used that year, whereas in 1926 every form used had been given in either 1924 or 1925. The effect of previous acquaintance with a particular test form would probably be felt most in the case of spelling and reading, and it is precisely on these two tests that the upturn is most pronounced.

INFLUENCE OF TIME UPON VARIABILITY

In Table XVII are assembled the standard deviations for all tests and types of score. Figures are given as after Sheppard's correction for effects of grouping.[16] The column at the right of the table shows the average of the standard deviations for the intelligence, spelling, reading, and Woody-McCall tests. As in Table XVI, geography is not included, owing to the scores in 1922 and 1926 representing the average made on two test forms, thus reducing the dispersion. The means of the last column are graphically depicted in Figure 5.

Table and graph unite in indicating that:

1. For all types of secondary score (T, B, and Q) where direct comparison is possible, there is a well-defined and fairly consistent tendency for dispersion to increase with time and schooling. This holds true for all tests save geography.

[15] For evidence as to the possibilities of improvement as the result of such encouragement, see:

Gilchrist, E.R. "Extent to which Praise and Reproof Affect a Pupil's Work." *School and Society*, IV: 872-74, 1916.

Hurlock, E.B. "Evaluation of Certain Incentives Used in School Work." *Journal of Educational Psychology*, XIII: 23-24, 1922.

Gates, G.S. and Rissland, L.Q.. "Effect of Encouragement and Discouragement upon Performance." *Journal of Educational Psychology*, XIV: 21-27, 1923.

[16] Yule, *op. cit.*, p. 212.

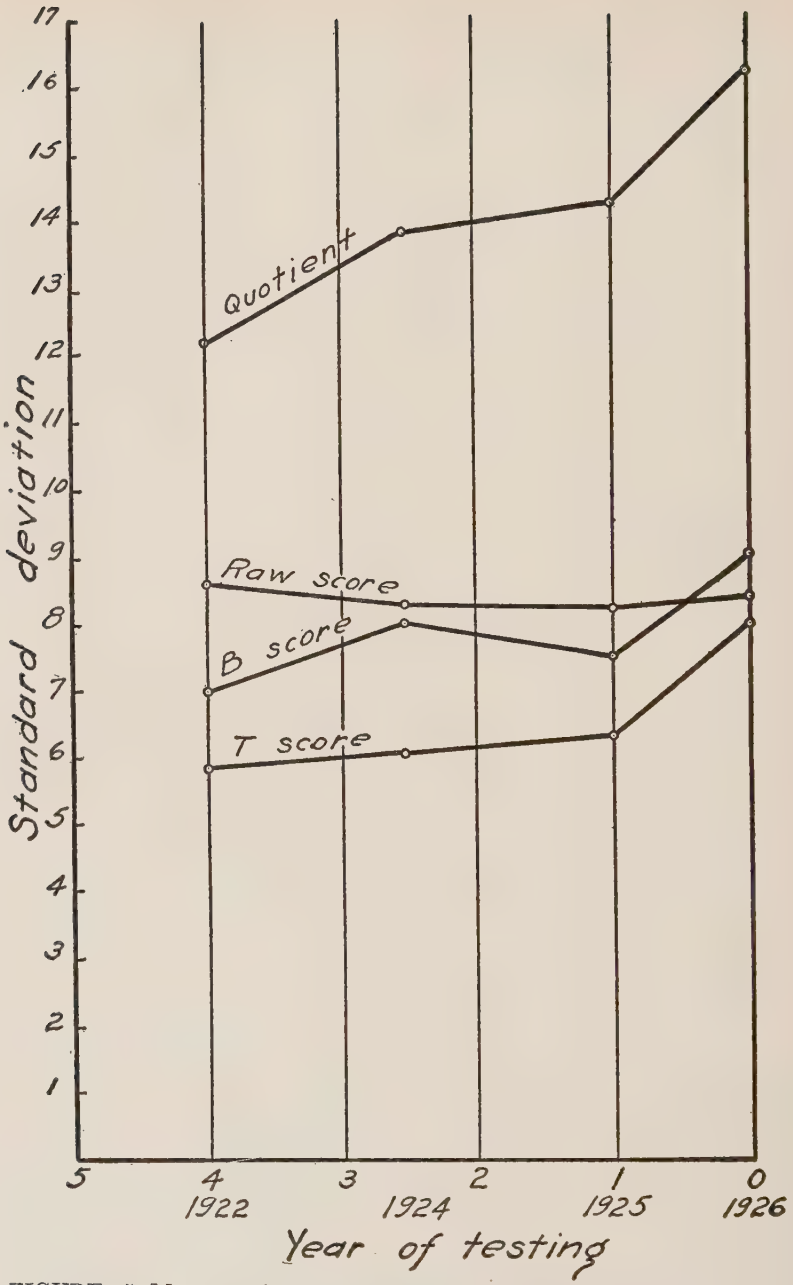


FIGURE 5. MEAN OF STANDARD DEVIATIONS OF CONSTANT GROUP OF INTELLIGENCE, SPELLING, READING, AND WOODY-MCCALL ARITHMETIC TESTS AT FOUR TIMES OF TESTING

Base line shows number of years prior to 1926, or criterion, measurement.

TABLE XVII

STANDARD DEVIATIONS OF SCORES OF THE CONSTANT GROUP, SHOWING TENDENCY TO INCREASE WITH SUCCESSIVE TESTINGS

Type of Score	Year of Test	Test							Mean S.D. apart from Otis and Geography
		N.I.T.–A	Spelling	Reading	Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean Educational Score	
Raw score on test	1922	18.7	6.77	4.31	1.50*	4.59	2.06	..	8.59
	1924	18.0	7.83	3.45	2.60	3.75	..	..	8.26
	1925	20.1	5.53†	3.07	2.95	3.23	2.68	..	8.23
	1926	18.1	8.49	3.38	2.73*	3.37	2.92	..	8.34
T score	1922	5.63	5.33	6.25	5.72*	6.13	6.10	4.20	5.84
	1924	5.35	5.03	6.44	8.02	7.43	..	..	6.06
	1925	5.68	5.94	6.60	7.13	7.08	6.96	4.28	6.33
	1926	6.32	9.74	7.71	6.99*	8.14	7.48	5.25	7.98
B score	1922	6.94	6.11	7.41	8.18*	7.62	7.60	6.03	7.02
	1924	6.78	6.50	7.74	10.27	8.87	..	..	7.97
	1925	6.87	7.69	7.59	8.26	7.66	7.93	5.95	7.45
	1926	7.33	11.30	8.54	7.87*	8.82	8.59	6.65	9.00
Quotient	1922	11.6	11.6	14.9	9.5*	10.7	14.1	10.00	12.2
	1924	11.4	12.1	15.9	18.3	15.7	..	..	13.8
	1925	13.5	12.6	15.9	15.7	14.7	15.6	11.07	14.2
	1926	15.2	14.5	18.3	15.4*	16.5	16.6	11.95	16.1

* Geography scores for 1922 and 1926 represent an average from two forms of the test, resulting in a marked reduction in variability.

† Number of words correct on the Morrison-McCall Spelling Test used in 1925 is not directly comparable with that on the Nifenecker Spelling Test used in other years.

Probable Error: As N is 200 in each case, the Standard Error of the above Standard Deviations is exactly five per cent, or one-twentieth, of the measure itself, and the Probable Error approximately one-thirtieth. Thus the Standard Deviation of spelling quotients for 1925 is 12.6 ± 0.63 S.D. error, or 12.6 ± 0.42 P.E.

2. In marked contrast to the above, the dispersion of raw scores is wholly erratic, depending on the units of the particular test concerned, and showing no common tendency either to increase or to lessen with time. Thus the change to Form 2 of National Intelligence Test A in 1925, as compared with Form 1 in other years, is marked by a sudden ten per cent increase in standard deviation of test scores.

3. The factors noted in the foregoing section as operating to reduce mean scores for 1925 and increase those for 1926 some-

what above expectation have seemingly had a like influence in the case of standard deviations.

For the benefit of those seeking further data on the effect of practice upon variability, Table XVIII is furnished. This shows

TABLE XVIII

RELATIVE VARIABILITY IN TERMS OF THE PEARSON COEFFICIENT ($V=100\sigma/M$) AS BASED ON T SCORES OF THE CONSTANT GROUP ON SUCCESSIVE TESTINGS

Test	1922	1924	1925	1926
Intelligence	15.3	11.6	10.8	11.2
Spelling	14.0	11.2	11.4	16.4
Reading	15.2	13.1	12.3	13.1
Woody-McCall	17.3	15.4	13.4	14.6
Geography	18.0	18.5	14.4	12.8
Otis	14.9	..	13.6	13.4
Mean educational score	11.4	..	8.2	9.2

the relative variability in terms of the Pearson Coefficient of Variation,[17] as based on T scores of the group on successive testings. The results agree with the preponderance of the findings assembled by Peterson and Barlow [18] in showing a progressive convergence, or reduction in variability with passage of time. The specious quality of comparisons made by any such formula when based upon units of uncertain equality and without reference to an absolute zero must, however, be emphasized. It has apparently escaped the authors of the article cited that this lies at the root of the unaccountable discrepancies and conflicting results observed.

[17] Yale, *op. cit.*, p. 149.

[18] Peterson, Joseph and Barlow, M.C. "The Effects of Practice on Individual Differences." In *Twenty-seventh Yearbook, National Society for the Study of Education*, Part II, Chap. XIV, especially p. 224.

CHAPTER IV

RESULTS: THE ADVANTAGES OF CUMULATION

The evidence adduced in Chapter III has shown more or less clearly that, under conditions of the present experiment:

1. Correlations between results secured with the same or parallel forms of a number of typical group tests show a clear and fairly regular decline with increase in time interval between testings.

2. The effects of such increase in interval are, when measured in terms of increase in the coefficient of alienation, practically identical for the four types of scores considered; whether these are expressed in terms of test units or of attainment with reference to age.

3. There is reason to believe that, up to a period of about three years at least, the relation between size of coefficient and length of interval is approximately rectilinear, and certain formulas have been developed by which to correct for the effects of time elapsed.

The present chapter will seek to determine whether the decline in correlations with time, as above noted, is such as to prevent the successful utilization of test results secured from one to three years previous, to improve upon the reliability of a current measurement, or if not, by what methods and with what degree of success such improvement may be had.

It is unnecessary to remind the statistically trained reader that the theoretical upper limit of a multiple correlation in which the correlation of every independent variable with the criterion and with every other independent variable may be assumed to be equal, is the square root of the zero-order coefficient.[1] The same might reasonably be regarded as representing the upper limit of a multiple correlation obtainable from an infinite num-

[1] Pearson, Karl. "On Certain Errors with Regard to Multiple Correlation." *Biometrika*, X: 181-87, 1914-15.
Kelley, T.L. "The Reliability of Test Scores." *Journal of Educational Research*, III: 327, 1921.

ber of test forms utilized to predict the score obtained from a single, criterion measurement such as 1926 in the present study, provided that all forms were so given as to eliminate the effects of practice, lapse of time, and the like. In other words, as the number of forms increases and the time intervals approach zero, the multiple R may be thought of as approaching $\sqrt{r_{11}}$ as a limit. The present study is not, however, concerned with the determination of such upper limits, which are already well known. It seeks, rather, to ascertain empirically the very limited amount of improvement to be expected in practice from the addition of

TABLE XIX

MULTIPLE CORRELATIONS BETWEEN ACTUAL 1926 PERFORMANCE AND PREDICTIONS BASED UPON TWO OR MORE PREVIOUS TESTINGS USING RAW SCORES IN THE ORIGINAL TEST UNITS

Correlation Coefficient	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
	N.I.T. –A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean T Score on Educational Tests*		
r_{12}	.807	.736	.614	.692	.618	.695	.749	.693	±.025
$R_{1(2+4)}$	.804	.771	.640	.719	.615	.721	.783	.710	±.024
$R_{1(24)}$	.825	.773	.663	.724	.656	.724	.792	.728	±.022
$R_{1(2+3)}$	.840	.784	.667	.750	.685	...	...	.745	±.021
$R_{1(23)}$	.841	.786	.676	.750	.687	...	...	.748	±.021
$R_{1(2+3+4)}$	.831	.781	.661	.757	.668	...	...	.740	±.022
$R_{1(234)}$	.846	.793	.683	.765	.697	...	...	.757	±.020

Subscript 1 denotes the 1926, or criterion, testing; 2 = 1925; 3 = 1924; 4 = 1922.

$R_{1(2+4)}$ indicates correlations resulting from a prediction based upon the simple (unweighted) average of 1925 and 1922 results.

$R_{1(24)}$ indicates the ordinary multiple correlation; i.e., the maximum obtainable by combining previous results, using the best possible two-place weightings. Etc., etc.

* In determining this average educational score of the individual, T scores were used, since the simple averaging of raw scores in the original units of the five educational, or achievement, tests would clearly be unsatisfactory.

results from a very few test forms given over long previous intervals, *the effects of time and practice upon correlations being what they are.*

THE MULTIPLE CORRELATIONS OBTAINED

It is evident that it is possible to compute from the data of Tables VIII to XI, the multiple correlations obtainable between the criterion measurement of 1926 and a combination of the scores secured from any two or more of the earlier testings. This has been done for raw scores in Table XIX, for quotients in Table XX, and for B scores in Table XXI. Separate computations for T scores are not shown, since they are closely similar to those for raw scores. On the first line of each table the zero-order coefficients found in Tables VIII to XI between the 1925 and 1926 measurements are shown as r_{12}. The third, fifth, and seventh lines give, respectively, the multiple correlations of 1922

TABLE XX

MULTIPLE CORRELATIONS BETWEEN ACTUAL 1926 PERFORMANCE AND PREDICTIONS BASED UPON TWO OR MORE PREVIOUS TESTINGS USING SCORES IN FORM OF QUOTIENTS

For meaning of subscripts see note to Table XIX.

Correlation Coefficient	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
	N.I.T. –A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean Quotient on Educational Tests		
r_{12}	.833	.782	.644	.754	.644	.758	.853	.731	±.022
$R_{1(2+4)}$	.849	.824	.666	.776	.687	.769	.878	.760	±.020
$R_{1(24)}$	.857	.824	.668	.780	.689	.780	.880	.764	±.020
$R_{1(2+3)}$	.873	.819	.678	.811	.678	...	.885	.772	±.019
$R_{1(23)}$	.873	.819	.681	.813	.684	...	.885	.774	±.019
$R_{1(2+3+4)}$	.871	.830	.676	.802	.690	...	.885	.774	±.019
$R_{1(234)}$	.877	.830	.685	.814	.701	...	.888	.781	±.019

and 1925 with 1926; 1924 and 1925 with 1926; and 1922, 1924, and 1925 with 1926.

TABLE XXI

MULTIPLE CORRELATIONS BETWEEN ACTUAL 1926 PERFORMANCE AND PREDICTIONS BASED UPON TWO OR MORE PREVIOUS TESTINGS USING BRIGHTNESS SCORES

For meaning of subscripts see note to Table XIX.

Correlation Coefficient	Test							Mean r of the Five Tests Other Than Otis	P.E. for a Single r of This Size
	N.I.T. –A	Spelling	Thorndike-McCall Reading	New York Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean B Score on Educational Tests		
r_{12}	.852	.797	.656	.763	.651	.758	.860	.744	±.021
$R_{1(2+4)}$	.859	.821	.684	.777	.693	.781	.881	.767	±.020
$R_{1(24)}$	.871	.822	.690	.794	.700	.789	.885	.775	±.019
$R_{1(2+3)}$	.885	.825	.701	.816	.709	...	.888	.787	±.018
$R_{1(23)}$	.887	.825	.702	.821	.712	...	.889	.789	±.018
$R_{1(2+3+4)}$	.878	.830	.700	.806	.712	...	.886	.785	±.018
$R_{1(234)}$	.889	.831	.709	.822	.722	...	.891	.795	±.017

In the computation of the multiple R's mentioned, the iteration technique of Kelley and Salisbury [2] was used to determine the optimum two-place weight, or β, to be given the scores from each year. These R's represent, therefore, the highest correlations obtainable with the use of the best possible two-place weightings.[3] As was to be expected, with such optimum weighting, the addition of results from a third year will be found in all cases to improve upon the coefficient for a two-year-composite. A later section will show that the addition of a fourth year likewise adds to the coefficient, though the improvement here is slight.

[2] Kelley, T.L. and Salisbury, F.S. "An Iteration Method for Determining Multiple Correlation Constants." *Journal of the American Statistical Association*, XXI: 282-92, Sept., 1926.

[3] In most cases the resulting coefficient is within .001 of that obtainable from weightings carried to an unlimited number of places.

CORRELATIONS WITH THE UNWEIGHTED MEAN

The citation of multiple R's as above described, though in accordance with common practice, is deceptive, if the impression is conveyed that such are the correlations to be expected in actual school practice. The reasons are:

1. Should the optimum weights as determined in the present study be published, their use by others, even under analogous circumstances, would be likely to yield somewhat varying results. In an unpublished Master's thesis of the University of Chicago, Miss Sauer has reported the results of the application of multiple regression weights secured by one investigator to data obtained by others under closely similar conditions. In the majority of cases such weightings proved to have little more than a chance validity, and yielded results showing slight improvement over equal weights for all.[4]

2. To be of general utility, instructions as to weighting should be in a form which admits of application to intervals which may differ from those of the original study. But prolonged examination of the optimum weights as here ascertained failed to disclose any general formula which would apply as well to the prediction of 1925 as of 1926 results. It is true that almost any plan which assigns a heavier weight to the more recent tests possesses certain merits, and β's proportional to the reciprocal of the square root of the time elapsing between the giving of the respective tests and the giving of the criterion to be predicted work reasonably well, but it is doubtful whether the added reliability is at all commensurate with the additional labor involved.

3. Office workers in public school systems are few, and demands on their time are heavy. The necessity for even a simple weighting of averages is likely to result in no averaging at all. Even though the time were available, it must be recognized that the process of weighting greatly enhances the liability to error in the hands of the average clerical assistant. To insist upon weighting for the sake of a relatively small gain in reliability is likely to prove straining at a gnat to swallow a camel.

Fortunately for our present purpose, when scores from a variety of sources are combined by a process of simple averaging, the influence of each of the components upon the final composite

[4] Compare the merits of weighting test scores in a composite, as summarized by Freeman in *Mental Tests*, pp. 272-75.

is in direct proportion to the standard deviation of its scores.[5] The tendency observed in Table XVII and in Figure 5 for standard deviations to increase with time and growth acts, therefore, of itself to give greater weight to later, or more recent, measurements. In order to determine the reliability of the prediction to be obtained from a simple, unweighted average of the results from two or more previous testings, it is necessary only to substitute for the optimum weightings of the Kelley-Salisbury procedure the standard deviations of the scores in the respective years. The computation of the multiple R's will then proceed as before.

The multiple correlations as thus determined for unweighted averages are shown in the second, fourth, and sixth lines of Tables XIV to XXI. It will be seen from these that:

1. For all tests and types of score, an unweighted average of the 1924 and 1925 results, $R_{1(2+3)}$, yields a correlation with 1926 which is practically equivalent to that obtainable with the best possible weightings, $R_{1(23)}$.

2. The advantages of a simple averaging which involves the 1922 results, dating from three years prior to 1925, are more doubtful. $R_{1(2+4)}$ approximates less closely to $R_{1(24)}$, and $R_{1(2+3+4)}$ falls short of $R_{1(234)}$. This is particularly noticeable in the case of raw scores, whose standard deviations, as shown in Table XVII, disclose no consistent tendency to increase with age and school progress.

LIMITS TO THE ADVANTAGES OF SIMPLE AVERAGING

The above observations raise the question as to how far into the past it is likely to prove profitable to draw upon test results in order to improve upon the reliability of a more recent measurement. From the last columns of Tables XIX to XXI we draw the mean correlations for the five tests for which complete returns are available, as shown in Table XXII:

From inspection of these mean correlations and those for the individual tests in the complete tables, it is evident:

1. That a simple averaging with 1925 of results from the 1924 testing, 1.5 years previous, yields correlations with 1926 which are in all cases marked improvements on the results of 1925 alone, and approximate closely those obtained with optimum weighting.

[5] Compare Garrett: *Statistics in Psychology and Education*, pp. 277-81.

2. That a simple averaging with 1925 of results from the 1922 testing 3.0 years previous leads, in the case of quotients and B scores, to a marked improvement upon the prediction from 1925 alone, but falls distinctly short of the correlations obtainable with optimum weighting; and that, in the case of raw scores, such averaging may actually detract from, rather than improve upon, the prediction based on 1925 alone (cf. figures for National Intelligence Test).

3. That the addition of 1922 results to an average of 1924 and 1925 proves, even in the case of quotients and B scores, to be as likely to detract from the prediction obtained by the latter as to improve upon it, while in the case of raw scores the coefficients are almost uniformly lower.

TABLE XXII

MEAN CORRELATIONS WITH 1926

Years Correlated with 1926	Raw Scores		Quotients		B Scores	
	Using Simple Average	Using Optimum Weights	Using Simple Average	Using Optimum Weights	Using Simple Average	Using Optimum Weights
1925 alone	.693		.731		.744	
1925, 1922	.710	.728	.760	.764	.767	.775
1925, 1924	.745	.748	.772	.774	.787	.789
1925, 1924, 1922 ...	.740	.757	.774	.781	.785	.795

From the above it would seem that under conditions of the present experiment the time limit beyond which it would be unwise to go in drawing on the results of past tests for averaging with current measurements will lie somewhere between 1.5 and 3.0 years prior to the latest testings, and probably in the neighborhood of 2.0 years.

A means of determining this more exactly is found in the mean correlations from four comparable tests as shown in Figure 3, and the corresponding standard deviations of Figure 5. By reading directly from the lines of Figure 3 as plotted, without attempt at smoothing, we obtain the following representative, or "type," correlations.

	Number of Years Between Tests				
	0	1	2	3	4
Raw score	.725	.679	.633	.590	.563
T score	.715	.670	.625	.584	.559
Quotient	.783	.747	.711	.678	.655
B score	.791	.757	.723	.690	.662

Similarly, from the lines of Figure 5 we read the typical standard deviations for the several types of score, as follows:

	Number of Years Prior to the Criterion Testing				
	0	1	2	3	4
Raw score	8.34	8.23	8.25	8.37	8.59
T score	7.98	6.33	6.15	5.99	5.84
Quotient	16.1	14.2	13.9	13.3	12.2
B score	9.00	7.45	7.79	7.66	7.02

By ignoring the tendency observed in Chapter III for later correlations to exceed the earlier as the standard deviations increase, and assuming that the correlations above will hold between any two tests of the stated number of years apart, we may now construct a table of coefficients of correlation for T scores on a hypothetical "type" test as follows:

Number of Years Prior to the Criterion Testing	Number of Years Prior to the Criterion Testing			
	1	2	3	4
0	.670	.625	.584	.559
1		.670	.625	.584
2			.670	.625
3				.670
Corresponding σ's for type test	6.33	6.15	5.99	5.84

Similar tables can be constructed for correlations for raw scores, quotients and B scores.

Given the data in this form, the computation of multiple correlations proceeds as in the case of an actual test. The results are assembled in Table XXIII. For the convenience of the reader, the coefficients of alienation corresponding to each of the r's are shown in the lower half of the table.

A comparison of Tables XXII and XXIII shows that the relative merits of cumulation by simple average and weighted mean agree closely by the two methods of approach. The absolute gains of the multiple R's over correlations with 1925 alone are throughout somewhat smaller in Table XXIII. This was to be expected from the mode of reckoning, which assumed that correlations with later years would be no higher than with earlier.

TABLE XXIII

MULTIPLE CORRELATIONS FOR A HYPOTHETICAL "TYPE TEST" GIVEN AT ONE-YEAR INTERVALS OVER A PERIOD OF FIVE YEARS

Subscripts indicate the number of years prior to criterion test.

Coefficient	Raw Scores		T Scores		Quotients		B Scores	
	Using Simple Mean	Using Optimum Weights	Using Simple Mean	Using Optimum Weights	Using Simple Mean	Using Optimum Weights	Using Simple Mean	Using Optimum Weights
r_{01}	.679		.670		.747		.757	
$R_{0(14)}$	.695	.714	.693	.701	.769	.774	.773	.781
$R_{0(12)}$	.716	.718	.709	.711	.780	.782	.789	.791
$R_{0(123)}$	.726	.730	.715	.721	.785	.790	.793	.799
$R_{0(1234)}$ ...	.700	.732	.715	.733	.785	.794	.792	.802
k_{01}	.734		.742		.665		.653	
$K_{0(14)}$	.719	.700	.721	.713	.639	.633	.634	.625
$K_{0(12)}$	.698	.696	.705	.703	.626	.623	.614	.612
$K_{0(123)}$	.688	.683	.699	.693	.619	.613	.609	.601
$K_{0(1234)}$ ...	.714	.681	.699	.680	.619	.608	.611	.597

The trend noted in Chapter III and the well-known tendency for increase in standard deviation to result in rising coefficients indicate that the figures of Table XXII give the truer picture in this respect.

As regards the limits to the advantages of simple averaging, results with unweighted means agree in reaching their maximum with the prediction achieved from an average of three successive annual tests. It is evident that the improvement resulting from averaging proceeds with rapidly diminishing returns from each additional earlier test. More specifically—

1. Averaging with results from a test of one or two years previous improves markedly upon the prediction obtainable from a current testing.

2. The addition of results from a third and earlier annual test to this combination is, on the whole, likely to improve slightly upon the prediction from the two tests, but the improvement is so small that the burden of proof might be upon the investigator to show that the reliability is not actually reduced.

3. The addition of results from a test given three years previous adds almost nothing to the reliability of the mean of the first two or three, even with the use of perfectly fitting weights; while with unweighted averages, indications are that the effect would be definitely detrimental.

4. The improvement in prediction resulting from cumulation, as measured by reduction in the coefficient of alienation (k) in approximately the same for all four types of score.[6]

AVERAGE ERRORS OF ESTIMATE

The unreliability of predictions from the usual regression equation is commonly expressed in terms of the standard error of estimate. This represents the standard deviation of differences between estimated scores and those actually obtained. But the formulas by which this is calculated[7] apply only when the customary regression coefficients, or ideal weights, have been used, and in the present instance these have been shown to be both impracticable and unprofitable. Happily, the theoretical constancy of B scores and quotients makes possible a measure of the reliability of prediction from any combination of scores which is both feasible and illuminating. It will be recalled that under each test on the individual record cards (cf. Figure 2) were recorded the mean of the pupil's B scores and the mean of his quotients for the years 1925 and 1922, 1925 and 1924, and 1925, 1924, and 1922. With these computed, it is possible to calculate directly the difference between each individual's score in 1926 and his average score on any two or three of the earlier tests. Now if we may assume the essential constancy of B scores and quotients for the group as a whole,[8] the average of such differences becomes by definition the average error of estimate for the unweighted mean concerned. Furthermore, since these differences may be expected to distribute normally, such average errors can be converted into standard errors and probable errors of estimate by use of the equations: [9]

$$\text{P. E. estimate} = .8453 \text{ average error of estimate.}$$
$$\sigma \text{ estimate} = 1.2533 \text{ average error of estimate.}$$

[6] Compare p. 38.

[7] Garrett, *op. cit.*, p. 243.

[8] This assumption does little violence to the facts when tests have been standardized on a strictly comparable group. See, for example, the behavior of N.I.T. quotients.

[9] Garrett, *op. cit.*, p. 85.

As a matter of fact, of course, neither B scores nor quotients do remain strictly constant, and the average differences found as above will reflect not only the variable errors of measurement due to unreliability of the test, but also any constant differences between mean scores on successive testings, due to practice effects, improvement of instruction, and the like. On spelling, for example, in contrast to the relative stability of the quotients on most tests, the mean quotients of our constant group showed a marked and continuous rise over the four-year period, as follows:

Year of Test	Mean Spelling Quotient for Constant Group
1922	94.3 ± 0.6
1924	95.3 ± 0.6
1925	96.8 ± 0.6
1926	99.6 ± 0.7

It is evident from this that the average difference between spelling quotients of the individual in 1925 and in 1926 will be equal to the mean error of estimate, due to unreliability of the tests, *plus* 2.8, or the difference between 1926 and 1925 performance of the group which is over and above the progress normally made in one year by the group on which the test was standardized. It is also evident that if quotients of 1922 are averaged with those of 1925, the mean will fall from 96.8, as in 1925, to 95.55, and the constant difference from 1926 will rise from 2.8 to 4.05. Unless, therefore, the averaging of 1922 results with 1925 has the effect of reducing the variable errors of measurement by more than 1.25, the amount which the constant error has been increased, the average of 1922 and 1925 quotients will agree less closely with 1926 than did the 1925 quotient originally.

The practical importance of this is manifest. An increase in the size of the correlation coefficient alone is no assurance of an improved prediction of quotients or B scores if a constant trend of the group, either for better or for worse, causes the averaging of results from earlier testings to introduce too large a factor of constant error. The figures of Table XXIV are, accordingly, of peculiar interest, as indicating the extent to which the mean errors of estimate were found actually to be reduced by simple averaging, under conditions approximating those to be expected in a typical school system, and without attempt to eliminate the influence of constant errors of any sort.

TABLE XXIV

MEAN ERROR OF ESTIMATE, OR AVERAGE DIFFERENCE BETWEEN THE STANDING OF INDIVIDUAL PUPILS IN 1926 AS MEASURED AND AS PREDICTED BY SIMPLE AVERAGING OF THEIR STANDING IN PREVIOUS YEARS

Years Compared with 1926	Intelligence	Spelling	Reading	Geography	Woody-McCall Arithmetic	Otis Arithmetic	Mean Educational Score	Mean of the 6 Columns apart from Otis
AVERAGE DEVIATION FROM 1926 QUOTIENTS								
1925 alone	6.34	8.15	11.50	8.73	10.27	8.65	5.27	8.38
	±.2	±.3	±.4	±.3	±.4	±.3	±.2	
1925, 1922	5.92	7.76	10.77	8.04	9.49	8.24	4.88	7.81
	±.2	±.3	±.4	±.3	±.3	±.3	±.2	
	6.6	*4.8*	*6.3*	*7.9*	*7.7*	*4.7*	*7.4*	*6.8*
1925, 1924	5.42	7.40	10.45	7.84	9.74		4.53	7.58
	±.2	±.3	±.4	±.3	±.3		±.2	
	14.5	*9.2*	*9.1*	*10.1*	*5.2*		*14.0*	*10.4*
1925, 1924, 1922 .	5.70	7.38	10.16	7.45	9.47		4.64	7.45
	±.2	±.2	±.3	±.3	±.3		±.2	
	10.1	*9.4*	*11.6*	*14.6*	*7.8*		*11.9*	*10.9*
AVERAGE DEVIATION FROM 1926 B SCORES								
1925 alone	2.93	6.41	5.67	4.60	5.47	4.46	2.85	4.66
	±.1	±.2	±.2	±.2	±.2	±.2	±.1	
1925, 1922	3.00	6.13	5.48	4.25	5.12	4.21	2.73	4.45
	±.1	±.2	±.2	±.1	±.2	±.1	±.1	
	−2.4	*4.4*	*3.4*	*7.6*	*6.4*	*4.5*	*4.2*	*3.9*
1925, 1924	2.60	6.02	5.03	4.16	5.15		2.47	4.23
	±.1	±.2	±.2	±.1	±.2		±.1	
	−11.2	*6.1*	*11.1*	*9.4*	*5.9*		*13.3*	*9.5*
1925, 1924, 1922 .	2.78	6.01	4.85	3.96	5.01		2.61	4.21
	±.1	±.2	±.2	±.1	±.2		±.1	
	5.1	*6.2*	*14.5*	*13.9*	*8.4*		*8.4*	*9.4*

Probable errors of the average differences in Table XXIV are shown to one place only, being in some cases estimates. Italicized figures beneath the various mean errors of estimate indicate the per cent by which the error shown is less than the corresponding error for 1925 alone. In the first column, for example, the figure *6.6* informs us that 5.92, which represents the mean error of estimate for intelligence quotients based on an average of 1922 and 1925 results, is just 6.6 per cent less than 6.34. This affords a ready indication of the extent to which the prediction obtained from the simple average has improved upon the prediction based on 1925 alone.

Examination of the data of Table XXIV, as summarized in these italicized figures, indicates that:

1. Mean errors of estimate, as thus empirically determined, agree closely with what would be expected from the behavior of the multiple correlations of Tables XXII and XXIII. The mean per cents of the final column show the general tendency for an average of 1922 results with 1925 to improve somewhat upon the prediction from 1925, for an average of 1924 and 1925 to improve still more, and for all three years combined to yield about the same reliability as 1924 and 1925 together.

2. This correspondence with Tables XXII and XXIII is most marked in the case of the quotients and B scores whose reliability is greatest, namely those for the intelligence and mean educational scores, and least in the case of Woody-McCall arithmetic, whose reliability by Table XV is lowest of the six tests.

3. This agreement is itself an evidence that there has been no general, constant trend for quotients and B scores to increase or decrease in proportions sufficient to outweigh the advantages of cumulation as indicated by the multiple correlation coefficients.

4. The averaging of 1922 results with 1925 tends on the whole to improve upon the latter, but the extent of this improvement is most uncertain, ranging from a gain of 7.6 per cent in the case of Woody-McCall quotients to an actual loss of 2.4 per cent for B scores in intelligence.

5. Predictions based on an average of 1924 results with 1925 represent a distinct improvement over predictions from 1925 alone, for all tests and types of score. The resulting reduction

in mean errors of estimate ranges from 5.2 per cent to 14.5 per cent and averages exactly 10 per cent.

6. The addition to the 1924-1925 combination, of results from a third test dating three years previous to 1925, adds, on average, little or nothing to the reliability of prediction from the two-year mean.

7. The precise amount of improvement in reliability resulting from simple averaging of B scores and quotients will vary considerably from test to test, depending not only on the correlation coefficients but on the constancy of mean scores. This suggests that for optimum results the individuals to be compared should have been tested with approximately the same frequency and at about the same times. This condition could best be met by a regular program of annual or semiannual tests given throughout a school or city system.

8. While mean differences in quotients and B scores on individual achievement tests (for example, 11.5 on Thorndike-McCall reading quotient, and 6.4 on spelling B score) are so great as to render these of little practical value, yet an average of five such fallible measures, as in the "mean educational score," is shown to possess a constancy over a year or more which is even greater than that of the National Intelligence Test A, and closely approaches those commonly reported for the Stanford Binet Test.[10] The implication is that by averaging the quotients or B scores obtained on tests of mediocre reliability in four or five subjects, given perhaps at somewhat different times during the term for instructional or other uses, an index of brightness may be obtained which is as reliable for purposes of next year's sectioning as that to be had from our best intelligence tests.[11]

COMPARATIVE ADVANTAGES OF DIFFERENT TYPES OF SCORE

In view of the size of the probable errors, as shown in Table XII, there can be no hope of demonstrating the greater reliability of one type of score over another. Data concerning the correlations to be obtained between T scores and B scores as compared with raw scores and quotients are, however, so rare in educational literature that it is well to observe the indications of the present study for what they may be worth.

[10] Compare Chapter I, references 5 to 13, and Freeman, *Mental Tests*, p. 355.
[11] Compare Gates and Lasalle, Chapter I, reference 21.

As between raw scores and T scores, the figures for the hypothetical "type" test of Table XXIII afford a convenient comparison. These correlations were based, it will be remembered, on the means of Tables VIII, IX, XIII, and XVII, and agree closely with the trend of Tables XIX to XXI.

The outstanding feature of these tables is the insignificant character of the differences. In some cases the raw scores yield higher correlations; in others, the T scores. Even if all questions of reliabilities be ignored and the mean differences taken at face value, the latter will be too small for practical significance.

The second and less certain conclusion is that nothing in these findings goes to point any advantage of T scores by the present mode of scaling over raw scores, on tests which have been standardized as carefully as these. Neither in absolute size of coefficients nor in their tendency to improve with cumulation do the T scores prove superior. In fact, the contrary is the case up to the addition of results from the third year previous. At that point the tendency for raw score units to prove unequal in value at different portions of the scale [12] appears to give the theoretically uniform T scores an advantage.

These two findings, coupled with the evidence that averaging of scores of more than three successive years adds nothing to reliability, would indicate that a general use of T scores in preference to raw scores must be justified on grounds other than reliability. Such grounds might be found in the convenient uniformity and ease of interpretation of T scores, the possibility of direct comparisons and averaging of test with test, and the simplicity with which locally developed and less carefully constructed test can be scaled by them.[13]

The evaluation of correlations for B scores and for quotients presents a nicer problem. It is clear that there is nothing to be gained by attempting a comparison between absolute scores, such as T scores and raw scores, and relative scores, such as B scores and quotients, since the functions and aims of the two types are wholly different. Nor would such a comparison be feasible. The so-called "spurious" factor [14] in correlations between in-

[12] See figures for the National Intelligence Test, p. 26.

[13] See also McCall, W.A. *How to Experiment in Education*, pp. 94-98.

[14] Yule, G.U. *An Introduction to the Theory of Statistics*, pp. 214-16. Thomson, Godfrey H. and Pintner, R. "Spurious Correlation and Relationship between Tests." *Journal of Educational Psychology*, XV: 433-44, October, 1924.

dices or ratios, such as are quotients, and between sums containing a constant term, as in the case of T scores, can be shown to yield a positive correlation approaching .50, even when no two of the variables are related. The same factors affect the comparison of correlations between B scores with those between quotients. In the absence of a statistical procedure which would be practicable in the present case for computing and eliminating those "spurious" elements, we must content ourselves with the figures as found. Fortunately, for our purposes this limitation is less serious than might be supposed. To quote from Yule,

> It does not follow of necessity that the correlations between indices or ratios are misleading. If the indices are uncorrelated, there will be a similar "spurious" correlation between the absolute measurements $z_1.x_3 = x_1$ and $z_2.x_3 = x_2$, and the answer to the question whether the correlation between absolute measures in misleading depends on the further question whether the indices or the absolute measures are the quantities directly determined by the causes under investigation.[15]

A like argument will apply in the case of B scores. Our present concern is not to compare the reliability of tests through the use of B scores and quotients, but rather to throw light on the question as to whether the ratio of performance to age or the sigma status of an individual within his age group remains the more constant with time.

To this question our data would suggest that, for the pupils of a typical elementary school grade over a period of four years, sigma standing as measured by B scores tends slightly to excel the quotient in reliability. As noted, however, in connection with Figure 3 and Table XIII, the differences in favor of B scores are too small to be either statistically reliable or practically significant.

In order to throw further light upon the relative merits of quotients and B scores, and the practical advantages of taking into account standing relative to age when making predictions over a period of time, the writer has computed the correlations obtained with and without the use of quotients and brightness scores. On the card shown in Figure 2, for example, the fourth space in the MA column indicates the MA predicted for 1926 on the assumption that the IQ of 1922 remains unchanged. As

[15] Yule, *op. cit.*, p. 216. Compare Yule, G.U. "On the Interpretation of Correlations between Indices or Ratios." *Journal of the Royal Statistical Society*, LXXIII: 644, 1910.

the interval was exactly 48 months, the prediction is easily obtained from the equation,

$$\text{MA } 1922 + \text{IQ } 1922 \times \text{interval} = 148 + 115 \times 48 = 203.$$

Similarly, the fourth space in the T score column contains the T score predicted for 1926 on the assumption that the B score of 1922 remains constant. The T score actually obtained in 1926 was 68, which corresponds to a B score of 60. In order for the B score to be 58 as in 1922, it is evident that John must have made a T score of only 66, as shown in space four. By finding the correlations between MA or T scores as thus predicted and those actually obtained, and comparing these correlations with those found between initial and final MA or T scores alone, we may discover whether or not the gains made have been in accordance with the initial brightness score obtained on the 1922 test.

Five such predictions were developed, using the comparatively reliable scores of the National Intelligence Test and the mean

TABLE XXV

CORRELATIONS BETWEEN FINAL SCORES AND THOSE PREDICTED WITH AND WITHOUT THE USE OF INITIAL MEASURES OF BRIGHTNESS (QUOTIENTS OR B SCORES)

	National Intelligence Test			Educational Battery		Mean Correlation
Years compared	1926, 1922	1925, 1922	1926, 1924	1926, 1922	1925, 1922	
Time interval (in months)	48	36	30	48	36	39.6
Correlation between:						
Final and initial age score	.650	.596	.755	.654	.582	.647
Final age score and prediction based on initial quotient	.669	.626	.780	.700	.632	.681
Initial and final quotients	.752	.741	.823	.814	.808	.788
Final and initial T scores	.650	.601	.743	.645	.587	.645
Final T score and prediction based on initial B score	.659	.656	.770	.710	.658	.691
Initial and final B scores	.750	.738	.835	.811	.805	.788

educational, or achievement, battery score. The results are given in Table XXV. In each case the prediction which takes into account the initial IQ or B score is found to improve somewhat upon the correlation between the initial and the final scores alone. The advantage is even greater in the case of the educational battery than with the test of general intelligence, so called. At best, however, the improvement is not great. Correlations for predictions based on quotients are somewhat larger than the correlations between age scores but much smaller than those between quotients; and the same is true with B scores. It is clear that the correlations of gains with initial measures of relative brightness are positive but not high. It must be remembered, also, that this finding is derived from measures of relatively high reliability and over periods of sufficient length to permit the rate of growth to offset the effects of chance errors of measurement. As the interval is shortened or the reliability of the tests lessened or both, it may confidently be anticipated that initial brightness will show a diminishing or negative relation to gains.[16]

Judged by the increased reliabilities of the predictions based on them as above, B scores show an average superiority to quotients about as in Tables XIII and XXIII. This is in accordance with logical expectation and the known deficiencies of the quotient technique as applied to point scales of this type.[17] The method of computing B scores, however, involves certain equally unwarranted assumptions. And, apart from all questions of statistical reliability, the significant feature of the differences above noted lies in their smallness. Taking the figures as they stand, it is clear that a superiority of .01 in correlations will be quite insufficient to weigh against considerations of practical convenience or expedience.

[16] Compare Thorndike, E.L. "The Influence of Chance Imperfections of Measures upon the Relation of Initial Scores to Gain or Loss." *Journal of Experimental Psychology*, VII: 225-32, June, 1924.

Thompson, Godfrey H. "A Formula to Correct for the Effect of Errors of Measurement on the Correlation of Initial Values with Gains." *Journal of Experimental Psychology*, VII: 321-24, August, 1924.

Burke, Mary M. "An Experimental Study of the Comparative Gains Made by Pupils Making High and Low Initial Scores." Master's thesis (unpublished). Syracuse University, 1927.

[17] For arguments as to the theoretical unsoundness of the quotient technique, see:

Rand, Gertrude. "A Discussion of the Quotient Method of Specifying Test Results." *Journal of Educational Psychology*, XVI: 599-618, 1925.

Freeman, *op. cit.*, pp. 275-85.

THE IMPROVEMENT EVALUATED

To permit of ready analysis and comparison, the more significant figures of Tables XIX to XXI are shown assembled in the first three columns of Table XXVI. These are:

r_{12}= the correlation obtained between the performance of individuals as measured in 1926 and in 1925. This represents the reliability of a prediction based upon results from a single form of the several tests concerned, given one year previously.

$R_{1(2+3)}$ = the correlation obtained between the measured status of individuals in 1926 and a simple unweighted average of their performance on the tests of 1925 and 1924, dated 1.0 and 2.5 years previous. This represents the combination found to yield an improved prediction with all tests and types of score, without the necessity for weighting.

$R_{1(234)}$ = the coefficient of multiple correlation obtained between the measured status of individuals in 1926 and that predicted by a combination of 1925, 1924, and 1922 measurements, using the best possible two-place weightings in each case. These figures are shown for the sake of comparison with $R_{1(2+3)}$ only, and not with the thought that such weighting would actually be employed in general school practice.

In the next three columns these correlations are shown as corrected for effects of the one-year interval between the 1925 and 1926 testing, through use of formula (3). They represent, therefore, the estimated reliability of the measurement, had the criterion test been given at the same time as the 1925, rather than one year later.

The last three columns give the ratio of the standard error of measurement to the standard deviation of the scores, and are thus equivalent to the radical $\sqrt{1-r_{11}}$. As explained by Kelley, $\sigma 1.\infty$ represents the proper value for the standard error of estimate when predictions are made as usual by taking the single score as an indication of true status in the trait measured.[18] Since the standard deviation of the distribution indicates the standard error of our estimate when we guess merely that an individual will be average for his group, and the standard error of measurement indicates the error of estimate when we infer that he will

[18] Kelley, T.L. *Interpretation of Educational Measurements*, pp. 171, 180-81.

hold the same relative position in his group as on previous tests, the ratio mentioned shows the per cent to which the error in a chance prediction is reduced through knowledge of previous standing of the individual. Thus the figure .407 in column 7 for raw scores on the National Intelligence Test indicates that, knowing the pupil's performance on Form 1 of the test, we can

TABLE XXVI

SUMMARY TABLE INDICATING DEGREE OF IMPROVEMENT ATTAINED THROUGH CUMULATION OF TEST RESULTS FOR VARIOUS TESTS AND TYPES OF SCORE WITH AND WITHOUT THE USE OF WEIGHTING

	Reliability of Prediction Over a One-Year Interval			Reliability of Measurement After Correction for Influence of Time Between Tests					
	r_{12}	$R_{1(2+3)}$	$R_{1(234)}$	r_{12}	$R_{1(2+3)}$	$R_{1(234)}$	$\frac{\sigma_1 \cdot \infty}{\sigma_1}$ for r_{12}	$\frac{\sigma_1 \cdot \infty}{\sigma_1}$ for $R_{1(2+3)}$	$\frac{\sigma_1 \cdot \infty}{\sigma_1}$ for $R_{1(234)}$
Raw scores									
Intelligence	.807	.840	.846	.834	.862	.868	.407	.372	.363
Spelling	.736	.784	.793	.773	.814	.822	.475	.431	.421
Reading	.614	.667	.683	.668	.714	.727	.580	.535	.522
Geography	.692	.750	.765	.735	.785	.798	.514	.464	.448
Arithmetic	.618	.685	.697	.671	.729	.739	.571	.521	.510
Mean r	.693	.745	.757	.736	.780	.791	.508	.465	.452
Quotients									
Intelligence	.833	.873	.877	.856	.890	.894	.379	.332	.325
Spelling	.782	.819	.830	.813	.844	.854	.432	.394	.381
Reading	.644	.678	.685	.694	.723	.729	.553	.526	.520
Geography	.754	.811	.814	.788	.837	.840	.460	.403	.400
Arithmetic	.644	.678	.701	.694	.723	.743	.553	.526	.506
Mean r	.731	.772	.781	.769	.803	.812	.475	.436	.426
Educational score	.853	.885	.888	.874	.901	.904	.355	.315	.309
B scores									
Intelligence	.852	.885	.889	.873	.901	.905	.356	.315	.308
Spelling	.797	.825	.831	.825	.850	.855	.418	.387	.380
Reading	.656	.701	.709	.704	.743	.750	.544	.507	.500
Geography	.763	.816	.822	.796	.842	.847	.451	.397	.391
Arithmetic	.651	.709	.722	.700	.750	.761	.548	.500	.488
Mean r	.744	.787	.795	.780	.817	.823	.463	.421	.413
Educational score	.860	.888	.891	.880	.904	.906	.346	.310	.306

predict his true standing in this type of intelligence with only 40.7 per cent as great an error as if we knew merely what was the average intelligence of the group from which he comes. And the adjacent figure .372 shows that this error will be further reduced to 37.2 per cent if prediction is based on average performance on the 1924 and 1925 tests. A comparison of the two columns will show that the standard error of estimate of the second is, on average, only about 91 per cent as great as that of the first. In this sense, then, the averaging in of scores from a test given 1.5 years previously is found to have improved about 9 per cent upon prediction from the most recent test alone.

To employ another criterion, Kelley has shown [19] that the difference between the means of a typical Low 6 and a typical High 6 grade is equal to $.52\sigma$ of the true ability scores of children in the combined sixth grade. The ratio of the standard error of estimate for any given r to this $.52\sigma_\infty$ may then be determined from the equation

$$.52\ \sigma_\infty = \mathrm{X}\ \sigma_1 \sqrt{1 - r_{1\mathrm{I}}}$$

in which X represents the number of standard errors of estimate. Or, since $\sigma_\infty = \sigma 1 \sqrt{r_{1\mathrm{I}}}$ by definition, we obtain the equation

$$\mathrm{X} = .52 \sqrt{\frac{r_{1\mathrm{I}}}{1 - r_{1\mathrm{I}}}}. \qquad (7)$$

Substituting for r 1I in the above, the value .834 from column 4, we obtain $\mathrm{X} = 1.1699$ standard errors. From the Kelley-Wood Tables of the probability integral, we then ascertain that this signifies that 75.8 per cent of the pupils will be measured correctly to the extent of distinguishing between differences in ability as great as that between the two successive half-year grades. By the same process, $R_{1(2+3)} = .862$ is found to yield $\mathrm{X} = 1.3042$ standard errors, and thus to indicate successful discrimination in 80.8 per cent of the cases.

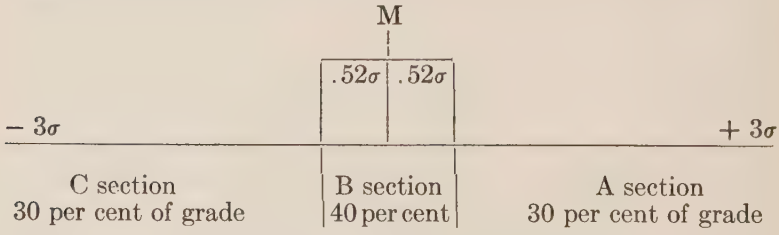
These results, which are fairly typical of the general table in the proportion of improvement shown, indicate that, whereas a single form of the test discriminates successfully between pupils of average High 6 grade and average Low 6 grade ability in 76 cases out of 100, an average of scores on two forms given one and one-half years apart distinguishes successfully in 81 cases

[19] Kelley, *op. cit.*, pp. 210-11.

out of 100. The improvement is not as great as we should wish. But it is far from insignificant for the five children formerly misplaced.

It will be evident, however, that the reliability as thus obtained with the National Intelligence Test must be regarded as minimal.[20] A test which does not discriminate successfully between differences of this magnitude at least four times out of five must be regarded as of little value for purposes of grade placement. The findings of this study would indicate that the only measures here examined which are likely to satisfy such a criterion are the National Intelligence Test and (by inference) the mean T score on the five educational tests combined. Even in these cases, an average would need to be taken of the results from two applications of the tests given not more than two years apart.

A similar analysis may be made of the significance of correlations in the case of quotients and B scores. Let us assume that the function of such measures will be to divide a typical year-grade into A, B, and C sections, including in the B, or average section, approximately 40 per cent of the enrollment, and in the A and C sections 30 per cent each. From the Kelley-Wood Tables we ascertain that in a normal distribution of ability, a distance of .5244 σ laid off on the two sides of the mean will include just 40 per cent of the population, as follows:



Substituting in equation (7) the value .5244 for the .52 as above, and taking $r_{II} = .856$, its value as between IQ's on two forms of the National Intelligence Test after correction for the

[20] For Kelley's own more exacting requirement, see pages cited. In this and the following discussion, however, it should be borne in mind that there is good reason to suppose from Table V that the constant group of the present study represents a selection less variable than that of a typical year-grade in a school system as a whole, although comparison with Miss Cronin's results would indicate that it approximates that of a year-grade class in a given school. If the definition of year-grade be taken to mean in a city system as a whole, the reliabilities would be considerably higher than here shown. It is unlikely, however, that this would seriously affect the major conclusions reached.

effects of time, we find X = 1.2785. This means that this difference in brightness will be distinguished successfully in 79.9 per cent of the cases. For $R_{1(2+3)}$ the value of X is 1.4914, indicating successful discrimination in 86.4 per cent of cases. These figures signify that with the use of quotients on a single form of the test the chances are that 1 in 5 of the pupils of average or below average brightness will be wrongly placed in the A section, and 1 in 5 of those of above average ability wrongly placed in the C section. By averaging in the quotient from another form of the test given 1.5 years before, the proportion thus misplaced is reduced to less than 1 in 7. For $R_{1(234)} = .894$, the per cent misplaced is 12.8 per cent, clearly but little improvement over the simple average above.

To take a further example, the mean B scores of the educational battery may be used, as representing the most reliable index of the present study. This gives for $r_{12} = .880$ just 15.6 per cent of the cases mismeasured as above, whereas by $R_{1(2+3)} = .904$ this is reduced to 10.8 per cent, and by $R_{1(234)} = .906$ to 10.4 per cent. As before, we find that the simple averaging of results from a single previous testing shows a notable improvement in the reliability, whereas the inclusion of a further and long previous test, even with ideal weighting, adds little to the foregoing.

CHAPTER V

SUMMARY AND CONCLUSIONS

It is common practice for teachers to give tests and examinations at various times during the year, and at the end of the term to average the marks made on these to determine promotions and failures. The teacher would be the last to deny that here and there a pupil may have been progressing at a rate which makes the averaging of earlier records a source of possible injustice. Given examinations which were entirely and unmistakably reliable, all questions of promotion or failure, classification, and the like would naturally be based on the latest measured status of the individual. But the best of such instruments are fallible. No one examination or test affords a completely trustworthy appraisal. By and large, the teacher believes, less injustice will result from decisions based on a number of measurements over a period of time rather than on any one, however recent.

The findings of the present study would imply that the teacher's judgment in this respect is sound. They go further and suggest that, even though the previous measurement whose results are averaged with the current one dates from as long as two years before, prediction of the pupil's next year's achievement will in all probability be improved—and materially improved—thereby.

This is a reasonable deduction from the correlations obtained as for simple averages of raw scores or T scores on the tests of the present study. To compare pupils directly on the basis of averages of raw or T scores made on tests given at widely different periods would, of course, be impracticable unless scores were available for all pupils to be compared, made on the same or parallel tests given at approximately the same time of year. Such a condition would be difficult to meet in practice; and in any case the average score would not lend itself readily to interpretation. If, however, the scores were cast into the form of percentiles, or preferably mean-square deviation status within

the grade, this difficulty would presumably be overcome. Resultant averages would then be meaningful even when standing alone, and pupil might be compared with pupil on the basis of these averages with little hesitation. This is what is involved in the computation of a multiple correlation; and this is what, in effect, teachers have long been doing by methods less precise. For teachers' marks are almost inevitably based on comparison, conscious or unconscious, with the stage of advancement of the group as a whole.

To reason from the results of the present study, an average of such grade-percentile or standard scores might be expected to fulfill the major administrative functions of our tests, so long as the present system of formal grades prevails. For pupils in Grades 5 to 8, at least, an average of several such grade-standing scores, dating from within the past two years, might be expected to serve the ends of promotion, sectioning, and the like from term to term and from year to year with very fair success. Assuredly the indications are that a prediction based on an average of two such scores from a test such as the National Intelligence Test would afford a more satisfactory forecast of next year's performance than would the results from a single form plus a knowledge of the pupil's IQ or Brightness score. In the reasonably well-graded school it would seem that it is only for the individual of outstanding brilliancy or stupidity, or for predictions looking some years into the future that the measure of brightness in relation to age assumes a major importance.

The present study is subject to several well-defined limitations:

1. It is based exclusively on the performance of pupils in Grades 4 to 8. At this age, however, growth in intellectual capacity and scholastic achievement is still progressing at a comparatively rapid rate. Any constancy of performance discovered at this stage would presumably be less than in later life, and Woodyard's observations as to the diminishing effects of time upon variability in the older groups tend to confirm this. The advantage which cumulation is found to possess may, therefore, be expected to hold good *a fortiori* for levels above Grade 8.

2. The tests used have, for the most part, been restandardized on the basis of norms prevailing in the schools from which our group is drawn. In the case of the one exception, spelling, the

behavior of test scores is closely similar to that of the rest. Conclusions are based, however, on the assumption of appropriate scaling, and may or may not hold good when the author's norms and scaling are accepted without attempt at adaptation to local standards and rates of progress.

3. The study deals only with group tests of the usual, point-scale varieties. Conclusions are not to be taken as applying to age-scales of the Binet-Simon type.

The major findings of the study, subject to the conditions named, are as follows:

1. That prediction of the performance of individual pupils one year in future, as based on the results of a single, current testing, is distinctly improved when the current findings are averaged with scores made on a parallel form of test given at any time within the past two years. In the present study this holds true without exception for all tests and types of score examined, and by all criteria available. The improvement resulting from a single averaging with the results of one other test dating 1.5 years previous was found to average approximately 9 per cent when measured by reduction in the standard error of estimate for different types of score, and 10 per cent when measured by reduction in the mean differences found between the quotients or B scores predicted and those which were actually obtained.

2. That the effect of diminishing returns is such that the addition of results from a further and a previous year's testing adds little to the reliability of a prediction based on an average of the last two or three annual tests, and may easily detract therefrom. The likelihood of detriment is naturally greatest in the case of raw or absolute scores, but is sufficiently serious even with quotients or B scores.

3. That the period in the past on which it appears we may profitably draw for such averaging with the expectation of worthwhile improvement and a minimum risk of detriment is approximately two years. Nothing in the present study would suggest that the averaging of results from any number of tests within this interval would have other than a desirable effect, provided the tests were not so distributed as to give greater weight to the earlier performance. In other words, we should expect a prediction based on a test given in May, 1927, to be

improved by averaging the 1927 scores with those made on parallel tests in May, 1926, and May, 1925, and any number of other tests evenly distributed over the two-year period.

4. That the weighting of scores utilized in such averaging is impractical and unnecessary, by far the greater part of the advantages of cumulation being had without this step. A tendency, noted in the case of T scores, B scores, and quotients, for standard deviations to increase with age and schooling acts in itself to give somewhat greater weight to more recent measurements.

5. That, although quotients and B scores on the separate achievement tests used prove too unreliable for individual prediction, a simple average of quotients or B scores from five such fallible tests possesses a constancy over one year which is greater than that of the National Intelligence Scale A, and comparable with those commonly reported for the Stanford Binet Test. The mean difference between quotients on the five-test educational battery of the present study in two successive years was found to be 5.27 ± 0.2; and the difference between an average of such quotients for two years and the quotients obtained in a third year was 4.53 ± 0.2. While these figures apply in strictness only to the particular combination of five tests here employed, it is reasonable to suppose that other combinations of group tests of like reliability and scaling would show similar results. If so, an average of quotients or B scores obtained from group tests in some half-dozen subjects, given perhaps at different times during the school year for instructional or other purposes, should yield a measure of brightness as useful for purposes of next year's sectioning as that obtained from one application of our best intelligence test.

6. That the improvement resulting from averaging is much the same for all types of score, and proportionally as great for the more reliable as for the less reliable tests.

7. Finally, it will be clear that the cumulation of test results has evolved no magic wand to make wholly inadequate tests adequate, or highly unreliable ones reliable. The limits to the advantages of averaging are fairly well defined and, in practice, somewhat narrower than might have been expected. The treatment is palliative rather than radical. Our findings indicate merely that by the simple and practicable process of unweighted

averaging of scores made at different times, the reliability of poor tests may be improved, and that of good tests made still better.

Out of the investigations incidental to the above, certain further observations and inferences have arisen:

1. For periods of one to four years, coefficients of correlation show a progressive lowering with increase in interval between tests. This influence of time is somewhat more pronounced than might have been inferred from the conclusions of Woodyard, based on intervals of one year or less. The relation in the case of these upper grade children is, however, so clearly marked over periods in excess of one year that it is improbable it should cease abruptly below that point.

2. The decrease in coefficients with lapse of time is found to be sufficiently regular and similar for all types of score to permit of an estimate of the reliability coefficient, when the correlation obtained over a period of years is known. While the evidence is inconclusive here, indications are strong that the relation between size of correlation and time between tests is not far from rectilinear, up to three years at least. On this assumption, certain formulas have been developed empirically by means of which coefficients of medium range obtained between tests given some time apart may be used to approximate the reliability coefficients for immediate retests.

3. Correlations between B scores show an average superiority of .013 ± .013 over correlations between quotients on tests standardized as were these. Such a difference is both statistically unreliable and practically negligible. Until some more satisfactory technique of scaling is developed, the choice of units for the indication of relative brightness will doubtless be left to considerations of practical convenience or expedience.

4. The improvement in prediction which results from taking into account brightness relative to age is not great, even with the more reliable instruments and over periods of three to four years. Over five intervals averaging forty months each, predictions taking into account the individual's quotient or B score on the initial test yielded an average correlation of .686 as compared with .646 for predictions based on T scores or age scores alone. As this was the difference found with the most reliable measures, the presumption is that the gain in the case of less reliable tests would be slight or negative.

5. For optimum results in prediction, rescaling of tests to conform to local norms and rates of progress is desirable. The differences between national norms as published by the various test authors and the norms actually prevailing in the trio of city schools concerned were such as to emphasize the importance of local restandardization of tests to be used in any widespread testing program.

6. Tests given under public school conditions by regular teachers trained for the purpose, to children scattered over more than thirty class sections in three widely separated schools, were found to yield correlations closely comparable with those obtained by experienced psychologists on retests of a single class or school. The evidence to this effect is one of the most encouraging outgrowths of the present study, for the promise which it holds of results to be anticipated from the widespread use of standard tests.

BIBLIOGRAPHY

Baldwin, B.T., and Stecher, L.I. "Additional Data from Consecutive Stanford-Binet Tests." *Journal of Educational Psychology,* XIII: 555-57, 1922.

Brooks, Fowler D. *Changes in Mental Traits with Age: Determined by Individual Retests.* Teachers College, Columbia University, Contributions to Education, No. 116. 1921.

Cronin, Mary Alice. "The Reliability of Certain Educational Tests." Master's Thesis (unpublished), Leland Stanford University. 1924.

Freeman, F.N. *Mental Tests.* Houghton Mifflin. 1926.

Garrett, H.E. *Statistics in Psychology and Education.* Longmans, Green. 1926.

Garrison, S.C. and Robinson, M.S. "A Study of Retests." *Journal of Educational Research,* XI: 190-96, 1926.

Gates, A.I. "The Unreliability of MA and IQ Based on Group Tests of Mental Ability." *Journal of Applied Psychology,* VII: 93-100, 1923.

Gates, A.I. and LaSalle, Jessie. "The Relative Predictive Value of Certain Intelligence and Educational Tests, Together with a Study of the Effect of Educational Achievement upon Intelligence Test Scores." *Journal of Educational Psychology,* XIV: 517-39, 1923.

Gates, A.I. and LaSalle, Jessie. "A Study of Writing Ability." *Journal of Educational Psychology,* XV: 205-16, 1924.

Jones, E.S. *Influence of Age and Experience on Correlations Concerned with Mental Tests.* University of Chicago Press. 1917.

Kelley, T.L. *Statistical Method.* Macmillan. 1923.

Kelley, T.L. *Interpretation of Educational Measurements.* World Book Co. 1927.

Kelley, T.L. and Salisbury, F.S. "An Iteration Method for Determining Multiple Correlation Constants." *Journal of the American Statistical Association,* XXI: 282-92, Sept., 1926.

McCall, W.A. *How to Measure in Education.* Macmillan. 1922.

McCall, W.A. *How to Experiment in Education.* Macmillan. 1923.

Peterson, J., and Barlow, M.C. "The Effects of Practice on Individual Differences." *Twenty-seventh Yearbook of National Society for the Study of Education,* Part II, Chap. XIV.

Pintner, R. "Group Tests after Several Years." *Journal of Educational Psychology,* XVI: 391-95, 1925.

Symonds, P.M. *Measurement in Secondary Education.* Macmillan. 1927.

Symonds, P.M. "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Journal of Educational Psychology,* XV: 423-32, 1924.

THORNDIKE, E.L. *The Measurement of Intelligence.* Teachers College, Columbia University, Bureau of Publications. 1927.

WOODYARD, ELLA. *The Effect of Time upon Variability.* Teachers College, Columbia University, Contributions to Education, No. 216. 1926.

LB
1131
K45

77294

JAMES C. KIRKPATRICK LIBRARY - UCM
3 7250 04235333 2
06-BEM-586